WORKING PAPERS

W.P. 84

ARCHIVIO DEGLI INDICATORI SOCIALI

Un approccio costruttivista all'organizzazione dei dati

Renato Miceli, Luca Ricolfi

ires

ISTITUTO RICERCHE ECONOMICO - SOCIALI DEL PIEMONTE

Working Paper n. 84

W.P. 84

ARCHIVIO DEGLI INDICATORI SOCIALI

Un approccio costruttivista all'organizzazione dei dati

Renato Miceli, Luca Ricolfi

Aprile 1988

I N D I C E

./.

NOTA INTRODUTTIVA

I tre studi che seguono sono il risultato di un lavoro di analisi teorica e di sperimentazione informatica condotto dall'IRES fra il 1986 e il 1987 nell'ambito delle attività dell'Osservatorio demografico.

Scopo di tale lavoro è la progettazione di un sistema di documentazione, archiviazione e accesso a dati di tipo socio-demogafico capace di combinare efficienza, flessibilità e facilità d'uso.

Il volume è articolato in tre parti: la prima è relativa al lavoro svolto nel primo anno; la seconda e la terza si riferiscono alle attività dell'anno successivo.

Nella prima parte -Progetto di fattibilità- vengono esposte nei dettagli le caratteristiche generali dell'architettura logica e funzionale dell'archivio degli indicatori sociali.

Nella seconda parte -Un approccio costruttivista all'organizzazione dei dati- viene resa esplicita la filosofia che ha orientato il disegno delle funzioni e delle prestazioni del sistema. Viene inoltre sviluppato e precisato un concetto -quello di "schema di disaggregazione"- che era rimasto in ombra nel progetto di fattibilità.

Nell'ultima parte -Il prototipo- vengono presentate le caratteristiche di una versione prototipale (già funzionante e utilizzabile) del sistema che si intende costruire. Tale versione consente nel giro di pochi minuti un accesso guidato ai dati nonchè la loro organizzazione automatica in matrici dati SAS (Statistical Analysis System) sulle quali l'utente può procedere a qualsiasi tipo di elaborazione statistica.

PARTE I PROGETTO DI FATTIBILITA'

1. INTRODUZIONE

1.1. Contenuto dell'Archivio

L'Archivio degli indicatori sociali dovrebbe contenere, a regime, da un minimo di qualche centinaio ad un massimo di qualche migliaio di indicatori di fenomeni di rilevante interesse sociale.

Le aree tematiche principali individuate finora sono dieci:

A - Comportamenti e strategie familiari
B - Reati
C - Comportamenti elettorali
D - Consumi culturali e istruzione
E - Scommesse
F - Comportamenti collettivi e associazionismo
G - Suicidio e altre cause di morte
H - Salute
I - Incidenti del traffico
L - Economia e mercato del lavoro.

L'ultima area (Economia e mercato del lavoro) non individua un insieme di indicatori propriamente sociali, ma è stata egualmente prevista per non far mancare all'Archivio un nucleo minimo di indicatori strutturali di sfondo, utili soprattutto in sede di analisi dei dati e di costruzione di modelli.

Ogni area è suddivisa in sottoaree, e ciascuna sottoarea comprende un certo numero di indicatori.

Idealmente un indicatore è una serie storica 1946-1986, dotata di una sua cadenza (mensile, trimestrale, annuale, ecc.) e disaggregata territorialmente (per regione, provincia, comune). Normalmente nell'Archivio vengono registrati soprattutto i dati che si riferiscono al Piemonte (regione, province, comuni) e, -per avere un termine di paragone- all'Italia nel suo insieme.

In pratica può accadere che le serie storiche inizino da una data più recente, e che il livello di disaggregazione territoriale effettivamente disponibile sia minore di quello ottimale.

Ogni indicatore registrato nell'archivio dà luogo ad una o più varianti, tante quanti sono i tratti non perfettamente confrontabili di una medesima serie storica (vedi appendice 4).

Una delle funzioni dell'archivio è appunto la "ricucitura" automatica (mediante procedure ad hoc) di varianti parziali, e quindi la produzione di serie storiche il più lunghe e omogenee possibile.

Ad ogni variante di un dato indicatore corrispondono una o più matrici dati rettangolari, ciascuna caratterizzata da una particolare organizzazione spazio-temporale (esempio: province piemontesi per trimestre; comuni piemontesi per anno ecc.). A loro volta ogni cella appartenente ad una di tali matrici è uno "slot" che può essere occupato da tre tipi di dati differenti, corrispondenti a tre diversi livelli di affidabilità:

I) Dato certamente definitivo
II) Dato potenzialmente definitivo
III) Dato certamente provvisorio.

Il significato esatto dei tre livelli di affidabilità è esposto analiticamente nel paragrafo 4.2..

La struttura logica dell'archivio può essere rappresentata nel modo seguente:

1. Area tematica
   2. Sottoarea
      3. Indicatore
         4. Variante
            5. Matrice spazio-temporale
               6. Cella della matrice
                  7. Tipo di dato (livello di affidabilità)
                     8. Dato vero e proprio.

Questa sequenza di "caduta" dal tema astratto al dato concreto può essere illustrata con un esempio.

1. Area tematica: comportamenti e strategie familiari
2. Sottoarea: nati legittimi e illegittimi
3. Indicatore: nati vivi maschi illegittimi riconosciuti
4. Variante: serie 1951-1986
5. Matrice spazio-temporale: Regioni per anni
6. Cella della matrice: Piemonte 1980
7. Tipo di dato: definitivo
8. Dato vero e proprio: 1070

La scelta degli indicatori da includere nell'Archivio dovrebbe ispirarsi a tre criteri principali.

Il primo è l'estensione temporale della serie che si può ottenere combinando tutte le varianti di un determinato indicatore. Fra gli scopi dell'Archivio vi è infatti quello di collocare i fenomeni analizzati in una dimensione storica, e di fornire a chi costruisce o usa modelli esplicativi o previsivi un materiale empirico di tipo longitudinale sufficientemente esteso.

Il secondo criterio è quello del livello di disaggregazione spaziale dell'indicatore. Idealmente l'Archivio dovrebbe essere in grado di fornire un quadro delle tendenze sociali e culturali ad un livello di dettaglio almeno provinciale.

Il terzo criterio è quello del contenuto di informazione. Un indicatore deve essere un "riassunto" il più possibile fedele di un complesso di indicatori di un dato fenomeno. Il contenuto di informazione di un indicatore può essere valutato quantitativamente con tecniche statistiche, rendendo così più agevole sia la scelta fra indicatori rivali sia il contenimento del grado di ridondanza dell'Archivio.

1.2. Scopi e funzioni dell'Archivio

L'Archivio degli indicatori sociali ha lo scopo di mettere a disposizione dell'utente un insieme di dati che è attualmente alquanto frammentario, disperso, eterogeneo, e risulta conseguentemente assai poco utilizzato. Soprattutto la mancanza di omogeneità rappresenta un grave ostacolo ad un pieno sfruttamento delle potenzialità interpretative implicite in molti indicatori sociali. Queste ultime si situano essenzialmente a tre livelli di analisi differenti.

A livello descrittivo l'Archivio dovrebbe consentire a chiunque lo desideri di conoscere l'andamento temporale e la distribuzione spaziale di qualsiasi fenomeno demografico, sociale o culturale. Da questo punto di vista il genere di domande cui l'Archivio dovrebbe saper rapidamente rispondere sono del tipo:

- quanti sono stati, nell'ultimo anno, i suicidi di minorenni in Piemonte?
- qual è la distribuzione territoriale, per provincia, dei conflitti di lavoro nell'ultimo anno?
- qual è stato l'andamento dei matrimoni nell'ultimo decennio nel comune di Torino?
- ecc..

A livello esplicativo l'Archivio dovrebbe consentire a chi desidera costruire un modello esplicativo di un determinato fenomeno di "saltare" tutta la complicata, noiosa e sovente delicata fase di predisposizione della base statistica.

A livello previsionale l'Archivio dovrebbe fornire sia la base statistica per costruire modelli orientati alla previsione e alla simulazione, sia alcuni strumenti standard "precotti" per effettuare previsioni, estrapolazioni e analisi di tendenza nei casi più semplici.

Questi scopi dell'Archivio ne richiedono un elevato grado di flessibilità, sia dal punto di vista delle modalità di interazione con l'utente, sia dal punto di vista delle prestazioni.

Sotto il primo profilo -quello delle modalità di interazione-

occorre dotare l'Archivio stesso di ampie capacità di comunicazione e dialogo in ciascuna delle tre fasi che caratterizzano l'interazione con l'utente:

a) alimentazione
b) informazione
c) analisi dei dati.

Sotto il secondo profilo -quello delle prestazioni- è opportuno che l'Archivio sia in grado di:

a) "ricucire" automaticamente serie discontinue o lacunose
b) costruire indicatori più o meno complessi a partire da serie grezze
c) fornire all'utente matrici dati in cui le variabili sono liberamente selezionate
d) trasformare in modo automatico una matrice dati in times series in una matrice dati in cross-section.

2. ARCHITETTURA LOGICA E FUNZIONALE

In questa prima parte del progetto l'architettura generale dell'Archivio verrà esposta esclusivamente da un punto di vista logico e funzionale, prescindendo dalle opzioni tecnico-informatiche connesse alla sua realizzazione concreta. Queste ultime riguardano innanzitutto la scelta fra sistemi locali, remoti e misti, e verranno trattate nella seconda parte del progetto.

2.1. Cinque tipi di data-base

L'Archivio degli indicatori sociali contiene cinque tipi di date base diversi:

1) un data base di dati originari (D1)
2) un data base degli indicatori (I)
3) un data base delle varianti (V)
4) un data base di dati derivati (D2)
5) un data base delle serie storiche derivate (S).

I primi tre (D1, I, V) si distinguono dagli ultimi due (D2, S) per i caratteri dei dati a cui si riferiscono. Nel primo caso i dati sono grezzi, o originari, corrispondono cioè esattamente ad una "pubblicazione" ufficiale (non importa se cartacea o magnetica) dell'ente preposto alla loro diffusione. Nel secondo caso costituiscono rielaborazioni più o meno spinte di dati del primo tipo.

All'interno dei due gruppi di data base abbiamo distinto fra data base di informazioni (I, V, S) e data base di dati (D1, D2).

	Serie originarie	Serie derivate
Informazioni	I V	S
Dati	D1	D2

Vediamo ora uno per uno i vari tipi di data base.

Nel data base I il record rappresenta un concetto astratto più o meno determinato (esempio: ore di sciopero nel terziario) di natura quantitativa (scala assoluta, di rapporti o di intervalli) e dotato di una o più definizioni operative. I campi del data base sono costituiti da tutti i descrittori del concetto stesso, e sono prevalentemente variabili di tipo carattere o di tipo qualitativo.

Nel data base V il record rappresenta una variante di un dato indicatore, e i campi sono costituiti, di nuovo, da un insieme di descrittori. Il data base contiene tutte le varianti di tutti gli indicatori contenuti nel data base degli indicatori, ed è collegato a quest'ultimo da una chiave che permette ad ogni variante di "puntare" all'indicatore da cui proviene, "ereditandone" così le caratteristiche.

Nel data base S il record rappresenta una particolare serie storica derivata, ossia costruita a partire dalle serie storiche originarie descritte in I e V. I campi sono costituiti dai descrittori della serie storica, che in questo caso altro non sono che la traccia delle operazioni attraverso cui è stata ricavata dalle serie storiche originarie da cui proviene.

Il data base D1 può essere visto come un insieme di matrici dati tridimensionali, che contengono esclusivamente numeri e possono assumere solo poche configurazioni standard (vedi paragrafo 4). Le tre dimensioni delle matrici sono lo spazio, il tempo e il livello di affidabilità del dato. Ogni matrice si riferisce ad una sepcifica variante di un dato indicatore. Una specifica variante può dar luogo a più di una matrice dati (questa eventualità si presenta ogni qualvolta una variante di un indicatore possiede più di uno schema di disaggregazione spaziale).

Il data base D2 ha la medesima struttura del data base D1 ma possiede una dimensione in meno (manca la dimensione livello di affidabilità).

Il cuore dell'Archivio degli indicatori sociali è evidentemente costituito dai data base D1 e D2, che contengono i dati veri e propri. Per usare l'Archivio, tuttavia, è indispensabile sapere che cosa c'è

in D1 e D2, e questa informazione si trova nei dati base I, V e S: le informazioni che costituiscono questi ultimi non sono altro che notizie sui dati contenuti in D1 e D2.

2.2. La distribuzione indicatore-variante

Che cosa distingue un indicatore da una variante?

Rimandando ad una appendice apposita una presentazione più sistematica della struttura dei due data base relativi ci limitiamo qui ad esporre il nucleo logico della distinzione. L'idea base è che, date due serie storiche concettualmente affini, internamente consistenti, dotate della medesima cadenza e del medesimo riferimento spaziale e tuttavia non coincidenti il passaggio dall'una all'altra possa essere rappresentato mediante una opportuna funzione di deformazione che trasforma la serie y nella serie y' e viceversa:

$$\begin{cases} y' = d(y) \\ \\ y = d^{-1}(y) \end{cases}$$

Poste le cose in questi termini ci si può chiedere se la forma della funzione d è tale da alterare in modo grave l'andamento temporale della serie di riferimento y. La risposta a questa domanda fornisce un primo criterio di distinzione tra indicatori e varianti: per essere considerate varianti del medesimo indicatore due serie storiche debbono essere legate da una funzione di deformazione semplice. Naturalmente questa risposta ha un senso ben definito solo se si riesce a precisare analiticamente che cosa si intende per funzione di deformazione semplice. Daremo quindi la seguente definizione:

Definizione. Una funzione di deformazione è semplice se equivale ad una trasformazione affine:

$$y = a\, y_t + b \qquad\qquad a > 0$$

Chiedersi se due serie storiche sono varianti del medesimo indicatore significa dunque chiedersi se è legittimo assumere che le differenze di andamento fra le due serie possano diventare trascurabili mediante semplici cambiamenti di unità di misura e di

origine.

Questo tipo di ragionamento si basa, a sua volta, su una classificazione più generale dei tipi di operazioni che possono essere compiute su una serie storica per trasformarla in un'altra e ci condurrà ad una caratterizzazione ulteriore della distinzione fra indicatori e varianti. La classificazione in questione considera quattro tipi di trasformazione di una serie:

1. Trasformazioni di scala (è il caso appena discusso)
2. Operazioni di aggregazione spaziale e temporale
3. Operazioni di disaggregazione spaziale e temporale (esempio: da dati annuali a dati trimestrali)
4. Trasformazioni complesse (stime di serie mancanti mediante loro proxies; costruzione di indici ecc.)

I primi tre tipi di trasformazioni, indipendentemente dalla facilità e dal rigore con cui possono essere realizzate nei casi concreti (la scala dei riferimenti temporali, ad esempio, è più facile da salire che da discendere), possono essere definite elementari, l'ultimo tipo di trasformazioni no. Solo le prime, infatti, possono essere realizzate con modelli statistici poveri di teoria e quindi in linea di principio completamente automatizzabili. Su questa base siamo in grado di formulare una prima caratterizzazione intuitiva della distinzione fra indicatori e varianti. Due serie storiche che hanno in comune tutti i descrittori che caratterizzano un indicatore (vedi Appendice 2) possono essere considerare varianti del medesimo indicatore se esiste un insieme di trasformazioni elementari che trasforma l'una nell'altra. In caso contrario occorrerà considerare le due serie storiche come due indicatori distinti.

La filosofia implicita in questa distinzione apparirà più chiara se proviamo ad immaginare come l'Archivio degli indicatori sociali funzionerà in concreto.

Un'utente si siede a terminale e chiede se nell'Archivio sono presenti serie storiche sugli scioperi industriali. Il sistema gli risponde che sì, esistono molti dati sugli scioperi ma disgraziatamente il medesimo indicatore (numero di scioperanti

nell'industria) è presente in numerose varianti, che differiscono fra loro per cadenza e definizione operativa (il concetto di industria si è leggermente modificato dopo l'adozione della classificazione SEC) e soprattutto coprono periodi storici disgiunti. A questo punto l'utente controlla che l'unione (in senso insiemistico) delle diverse varianti disponibili copra un periodo sufficientemente lungo. Se questo controllo ha esito positivo l'utente seleziona una variante (esempio: scioperi industriali in media annua, definizione SEC) e chiede al sistema di ricondurre tutte le altre varianti a quella desiderta utilizzando esclusivamente trasformazioni elementari. E' chiaro che questa operazione è possibile solo in quanto le serie storiche che vengono raccordate, o rese uniformi nella cadenza, differiscono fra loro per aspetti relativamente "triviali". Se il contenuto concettuale di due o più serie fosse sostanzialmente differente non potremmo ricondurre l'una all'altra senza introdurre ipotesi e informazioni addizionali, senza formulare cioè una teoria specifica sui rapporti fra le serie considerate.

Il nucleo concettuale della distinzione fra indicatori e varianti è appunto questo: due serie vanno trattate come indicatori distinti anzichè come varianti del medesimo indicatore se il passaggio dall'una all'altra non può essere effettuato mediante modelli poveri di teoria.

2.3. Serie storiche originarie e derivate

Accanto ai data base degli indicatori e delle varianti originari (I e V) esiste una data base ausiliario (S) che contiene le descrizioni di tutte le serie storiche construite a partire da serie storiche originarie. Le serie storiche derivate sono essenzialmente di tre tipi:

a) serie ricavate da serie originarie mediante trasformazioni elementari di varianti (raccordi, cambiamenti di cadenza, ecc.)
b) serie ricavate da serie originarie per aggregazione di serie elementari (esempio: popolazione totale = minorenni + maggiorenni)
c) serie ricavate da altre serie attraverso operazioni di confronto o di sintesi fra indicatori differenti (calcolo di rapporti, differenze, medie, ecc.).

L'esistenza di due differenti tipi di data base (serie originarie e derivate) ha una triplice funzione.

In primo luogo essa ha il compito di porre dei limiti precisi al potenziale di esplosione combinatoria inevitabilmente connesso ad un archivio di indicatori. Dato un fenomeno concreto (esempio: delitti) classificato in K classi il suo potenziale di esplosione combinatoria P è governato dalla relazione:

$$P = 2^K - 1$$

che si ricava immediatamente dal teorema dell'insieme delle parti escludendo l'insieme vuoto.

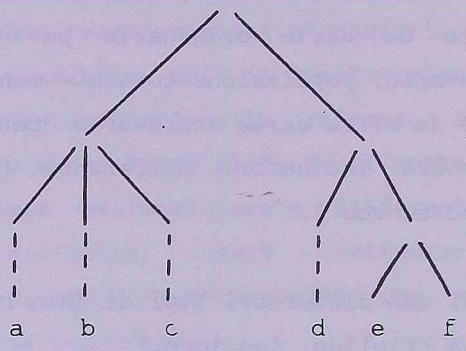
Questo significa, ad esempio, che con 10 serie elementari si possono generare fino a 1.023 ($1.023 = 2^{10}-1$) serie derivate.

Una eliminazione completa della ridondanza si potrebbe ottenere imponendo alle serie storiche descritte nei data base I e V di non essere ricavabili l'una dall'altra utilizzando esclusivamente le operazioni di somma e di differenza fra serie storiche. Questa condizione, tuttavia, pur essendo del tutto soddisfacente sotto un profilo logico, è alquanto difficile da rispettare sul piano pratico.

La sostituiremo quindi con un insieme di regole più deboli ma più facili da applicare.

Regola 1 Dato un insieme di serie storiche generate da una classificazione gerarchica registrare solo le serie elementari, corrispondenti ai livelli più fini della classificazione.

Esempio:



Si registrano solo le "foglie" finali a-f, trascurando tutti i nodi intermedi.

Regola 2 Dato un insieme di serie storiche generate da due o più classificazioni incrociate registrare solo le serie storiche elementari poste "all'intersezione" fra le classificazioni ed omettere quelle "marginali".

Esempio:

	maschi	femmine	
maggiorenni	a	b	r1
minorenni	c	d	r2
	c1	c2	TOT

Si registrano le serie elementari a, b, c, d trascurando i marginali di riga (r1, r2) e di colonna (c1, c2).

Regola 3 Dati due insiemi di serie storiche generati da due classificazioni indipendenti (non incrociate) registrare le serie elementari di entrambe anche se questo comporta una lieve ridondanza.

	maschi	femmine	
maggiorenni	?	?	r1
minorenni	?	?	r2
	c1	c2	TOT

Si registrano le quattro serie elementari r1, r2, c1, c2 anche se l'insieme considerato possiede solo tre gradi di libertà, e potrebbe quindi essere descritto esaustivamente utilizzando soltanto tre serie.

In secondo luogo la distinzione fra i due tipi di data base ha il ruolo di definire un confine molto netto fra dati e costrutti dell'utente, e soprattutto di consentire di ricostruire il modo in cui i secondi sono stati generati a partire dai primi.

Questo significa che nel data base S la descrizione di una particolare serie storica derivata consisterà semplicemente in un elenco ordinato di tutte le operazioni che hanno permesso di costruirla. Detto in altre parole: qualsiasi serie storica descritta in S deve essere ricavabile da serie storiche descritte in I e V mediante un insieme ben definito di operazioni, elementari e non.

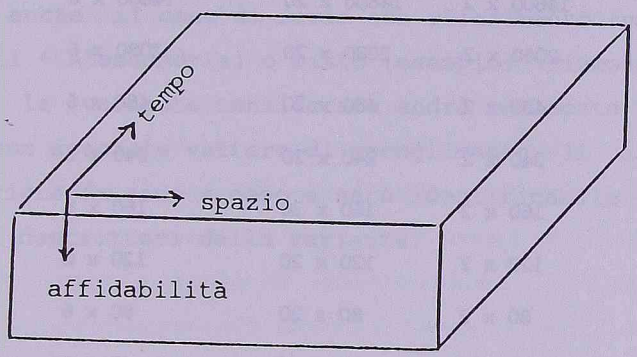
La distinzione fra i due tipi di data base, infine, ha un ruolo

pragmatico, nel senso che permette un accesso differenziato all'Archivio da parte di due fasce di utenti che hanno esigenze sostanzialmente differenti. Da un lato esistono utenti che desiderano esercitare un controllo completo e rigoroso sulle fonti delle proprie elaborazioni, e hanno quindi bisogno di accedere ai data base delle serie originarie o per costruire in proprio nuove serie, o per controllare il rapporto esistente fra determinate serie "costruite" e le rispettive serie sorgente. Dall'altra esistono utenti che, o perchè accettano il lavoro già fatto da altri o perchè hanno già effettuato i loro controlli in precedenza, preferiscono lavorare su serie "elaborate" (derivate) piuttosto che su serie "grezze" (originarie).

2.4. Matrici dati

Nei due paragrafi precedenti abbiamo parlato di tre data base (I, V, S) che contengono informazioni sul contenuto di altri data base (D1 e D2) nei quali risiedono i dati veri e propri. In questo paragrafo analizzeremo la struttura di questi ultimi tipi di data base.

Il data base D1 può essere visto come un insieme di matrici dati a tre dimensioni (spazio, tempo e livello di affidabilità), ognuna delle quali si riferisce ad una specifica variante di un determinato indicatore.



Il data base D2 ha una struttura analoga a quella del data base D1 salvo il fatto che le matrici dati hanno una dimensione in meno (manca, per ragioni che verranno esposte più avanti, la dimensione "livello di affidabilità").

2.4.1. Organizzazione spazio-temporale

Tutte le varianti descritte nel data base V possiedono una cadenza caratteristica. Inoltre possono essere disponibili su base nazionale, regionale, provinciale o comunale. Poichè le cadenze rilevanti sono 10

e i livelli di disaggregazione spaziale rilevanti sono 4 è possibile definire a priori un piccolo numero di schemi di organizzazione spazio-temporale -o moduli ordinatori- in cui far rientrare i dati relativi a qualsiasi variante considerata. Tali schemi sono in numero di 40 (10 x 4) e si ottengono incrociando la dimensione "cadenza" (tempo) con la dimensione "livello di disaggregazione territoriale" (spazio).

SPAZIO / TEMPO	Nazionale	Regionale	Provinciale (solo Piemonte)	Comunale (solo Piemonte)
giorno	14600 x 2	14600 x 20	14600 x 6	14600 x 1209
settimana	2080 x 2	2080 x 20	2080 x 6	2080 x 1209
mese	480 x 2	480 x 20	480 x 6	480 x 1209
bimestre	240 x 2	240 x 20	240 x 6	240 x 1209
trimestre	160 x 2	160 x 20	160 x 6	160 x 1209
quadrimestre	120 x 2	120 x 20	120 x 6	120 x 1209
semestre	80 x 2	80 x 20	80 x 6	80 x 1209
anno	40 x 2	40 x 20	40 x 6	40 x 1209
decennio	4 x 2	4 x 20	4 x 6	4 x 1209
irregolare	n x 2	n x 20	n x 6	n x 1209

In ogni cella sono riportate le dimensioni massime delle matrici dati associate ad ogni combinazione spazio-temporale. Come si vede solo nel caso dell'ultima riga della tabella (cadenza irregolare) la dimensione delle matrici dati è relativamente indeterminata.

Questa tipologia di schemi a priori consente di attribuire in modo univoco ad ogni variante un insieme di "matrici di accoglimento" la cui struttura non richiede di essere riconfigurata volta per volta.

Nel caso delle varianti descritte nel data base S, quello delle serie storiche derivate, nulla cambia salvo il fatto che la tipologia delle matrici di accoglimento acquista una colonna in più. Accanto alle serie dotate di un preciso riferimento spaziale occorre infatti considerare anche il caso di serie con riferimento spaziale multiplo (esempio: Asti + Alessandria) o misto (esempio: Piemonte/Italia). In questi casi la variante considerata andrà collocata -anzichè in una matrice- in uno speciale vettore di accoglimento, il cui significato spaziale varierà da caso a caso e sarà identificabile in modo univoco a partire dai descrittori della variante.

2.4.2. Livelli di affidabilità

La maggior parte dei dati pubblicati dall'ISTAT e dagli altri enti che producono in modo regolare informazione statistica (Ministeri, INPS, ecc.) subiscono nel corso del tempo una serie di modificazioni. E' abbastanza raro che alla prima "edizione" di un certo dato seguano "edizioni" identiche del medesimo dato. Di norma il dato stesso si stabilizza soltanto qualche anno dopo la sua prima apparizione.

Questo processo di aggiustamento può dipendere da vari fattori, alcuni dei quali strettamente inerenti al processo di costruzione del dato. Fra questi ultimi è il caso di ricordarne tre.

Il primo è la lentezza delle procedure amministrative di alimentazione dell'Istituto centrale di statistica da parte degli enti periferici (anagrafi, uffici di collocamento, tribunali, ecc.). Questa lentezza costringe sovente l'Istat a pubblicare statistiche basate su

rilevazioni incomplete, e a rivedere i dati iniziali man mano che gli enti periferici "ritardatari" inviano i rispettivi moduli.

Il secondo fattore di instabilità è costituito dal processo di pulizia dei dati che l'Istat è costretto a compiere per eliminare incongruenze ed errori di rilevazione. Tale processo può essere alquanto complesso, soprattutto quando non dipende da un unico insieme di dati, o quando gli insiemi di dati da cui dipende pervengono in tempi successivi.

Il terzo fattore di instabilità è costituito dal carattere di stime (anzichè di rilevazioni esaustive) di molti dati pubblicati dall'Istat. Gli esempi più importanti al riguardo sono quelli della contabilità nazionale e dell'indagine trimestrale sulle forze di lavoro. Questa fonte di instabilità è particolarmente grave nei casi in cui sono incerti addirittura i coefficienti di riporto all'universo. E' il caso di tutte le stime basate su un campionamento stratificato della popolazione. L'estrazione del campione richiede infatti conoscenze che divengono note solo al momento del censimento, ossia ogni 10 anni. In mancanza di tali conoscenze l'istituto centrale di statistica lavora su coefficienti di riporto stimati, e perfeziona le stime man mano che si modificano le sue congetture sulla struttura dell'universo. Questo processo di modificazione subisce naturalmente un'impennata in occasione di ogni censimento, e può coinvolgere tutte le stime del decennio che lo precede.

L'instabilità dei dati ha conseguenze piuttosto gravi sulla omogeneità delle serie storiche costruite a partire da essi. Il processo di aggiustamento, infatti, non ha una struttura statistica completamente random. In molti casi, ad esempio, i primi dati sono tendenzialmente sottostimati. Ciò significa che chi confrontasse un dato appena pubblicato relativo all'anno t con il dato corrispondente all'anno t-1 ma "ritoccato" (seconda pubblicazione) potrebbe farsi un'idea compltamente errata (per difetto) del saggio di variazione del fenomeno considerato.

Queste ed altre considerazioni simili hanno suggerito di conferire alle matrici del data base D1 (quello dei dati originari) una terza

dimensione, che rappresenta il livello di affidabilità del dato. A questo scopo sono stati identificati tre diversi livelli di affidabilità:

I. dati (sicuramente) definitivi
II. dati semidefinitivi o potenzialmente definitivi
III. dati (sicuramente) provvisori.

Il livello II è residuale nel senso che include tutti i dati che non possono esere attribuiti con sicurezza nè al livello più alto (dati definitivi) nè al livello più basso (dati provvisori). Restano dunque da definire in modo preciso i due livelli estremi.

Dati provvisori. Sono considerati provvisori tutti i dati che posseggono le due proprietà seguenti:

a) rappresentano la "prima edizione" del dato, o compaiono in una pubblicazione contemporanea alla "prima edizione" (per contemporanea si intende apparsa nel medesimo anno)
b) non vengono sistematicamente confermati nelle "edizioni" successive.

Dati definitivi. Sono considerati definitivi tutti i dati che posseggono le due proprietà seguenti:

a) sono congruenti con una o più serie storiche apparse nell'ultimo Sommario di statistiche storiche dell'Italia dell'Istat
b) hanno un indice temporle non successivo all'ultimo censimento considerato nell'ultimo Sommario di statistiche storiche dell'Italia.

Un'ultima considerazione sul funzionamento e sull'uso della terza dimensione delle matrici dati. Contrariamente a quanto si potrebbe pensare l'arrivo di versioni successive del medesimo dato, fino alla versione definitiva, non rende perciò stesso superflue le versioni dei livelli più bassi. Queste ultime devono essere mantenute perchè,

nonostante la loro minore vicinanza al dato definitivo, possono risultare utili nei casi in cui una certa serie storica non è interamente costituita da dati definitivi. In circostanze del genere la ricostruzione di una serie il più omogenea possibile richiede che, almeno negli anni di raccordo fra tratti caratterizzati da livelli differenti di affidabilità, si disponga di più di una versione del medesimo dato.

2.5. Interazione con l'utente

In quel che segue le tre funzioni base dell'Archivio -alimentazione, informazione e analisi dei dati- verranno descritte essenzialmente dal punto di vista dell'utente, senza entrare ancora nel merito dei modi di realizzarle sotto il profilo informatico.

2.5.1. Alimentazione

La funzione di alimentazione dell'Archivio si suddivide in due sottofunzioni principali:

1. inizializzazione di indicatori, varianti e serie storiche derivate
2. input di dati.

Si tratta di due funzioni profondamente differenti che richiedono competenze e abilità diverse, e saranno svolte -tendenzialmente- da persone differenti.

L'inizializzazione di indicatori e varianti consiste nel descrivere in modo accurato, rigoroso e completo le caratteristiche di un certo indicatore e di tutte le sue varianti (vedi Appendici 2 e 3). Questa operazione vale sia per le serie storiche originarie (indicatori e varianti) sia per le serie storiche derivate, anche se si svolge in modi sostanzialmente diversi per le une e per le altre. Mentre nel primo caso si tratta di "istanziare" una serie di descrittori predefiniti dell'indicatore e della variante, nel secondo si tratta di specificare in modo completo la catena di operazioni che conducono da un certo insieme di serie storiche originarie (contenute in D1) ad una serie storica derivata (contenuta in D2).

E' chiaro che in entrambi i casi le competenze richieste sono piuttosto elevate. Nel primo caso è indispensabile una certa conoscenza del dominio cui gli indicatori appartengono, e soprattutto una grande dimestichezza con le fonti statistiche da cui provengono i dati . Nel secondo caso è indispensabile soprattutto una buona

conoscenza statistica e metodologica delle procedure di costruzione degli indicatori, di raccordo fra serie storiche, di trattamento delle serie temporali e così via.

L'input di dati consiste invece nel riempire le matrici dati "di accoglimento" relative a una certa varianate di un certo indicatore già inizializzati. Questa operazione di ingresso-dati dovrebbe poter essere svolta secondo modalità relativamente elastiche, capaci cioè di adattarsi alla grande varietà di configurazioni in cui i dati da digitare si possono trovare. Più precisamente il sistema di ingresso dei dati deve essere in grado:

a) di leggere da dischetto matrici dati organizzate secondo il classico schema case by variable
b) di ricevere da un utente che li legge su carta dati organizzati in tabella.

Nel caso a) l'unica limitazione consiste nel fatto che il record deve essere o un'unità temporale (mese, trimestre, ecc.) o un'unità spaziale (regione, provincia, ecc.).

Nel caso b) l'unica limitazione è data dalla dimensione dello schermo del terminale. Poichè la struttura della tabella da copiare viene riprodotta sullo schermo mediante una sorta di "mascherina" flessibile diventa indispensabile fissare dei limiti alle dimensioni dei blocchi di tabella simulabili a video.

2.5.2. Informazione

Non sempre l'utente che si accosta all'Archivio degli indicatori sociali ne conosce con precisione il contenuto. Prima di utilizzare i dati memorizzati nell'Archivio può essere indispensabile capire che cosa c'è dentro l'Archivio. A questo scopo l'utente ha a disposizione un vero e proprio ambiente di ricerca, che gli consente di capire in pochissimi passi qual è la struttura e il contenuto dell'Archivio.

L'ambiente di ricerca ha una struttura alquanto complessa perchè deve essere in grado di fronteggiare una gamma molto vasta di

possibili richieste.

La prima differenziazione di cui tenere conto è quella fra i due data base "bersaglio", quello delle serie originarie (D1) e quello delle serie derivate (D2).

La seconda differenziazione è quella fra richieste temporalmente o spazialmente delimitate e richieste generalizzate, che non pongono alcun vincolo all'ambito di riferimento degli indicatori.

La terza differenziazione è quella fra utenti che hanno richieste generiche (che cosa c'è nell'Archivio ?) e utenti che hanno richieste specifiche, limitate cioè ad un singolo tema, o addirittura ad un singolo indicatore o ad un singolo dato.

La quarta differenziazione è interna alle richieste specifiche, e può essere caratterizzata mediante tre domande tipo:

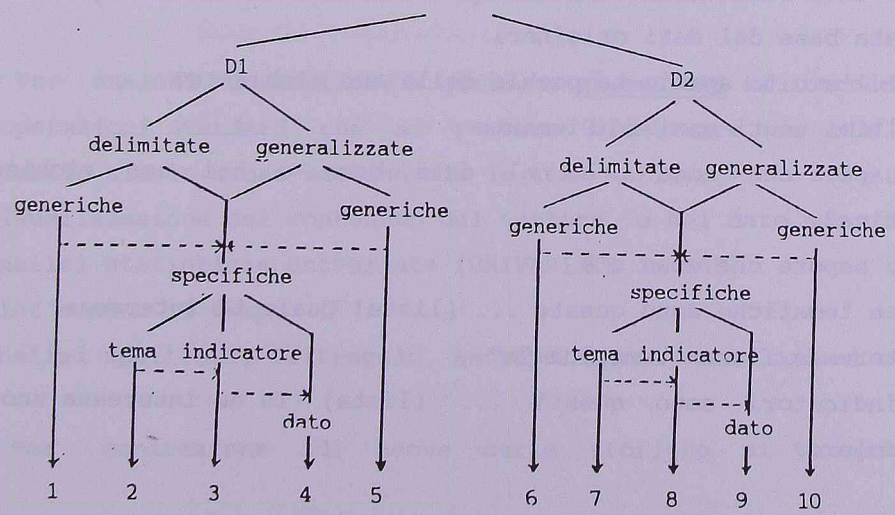
a) quali serie storiche sono disponibili sul tema "x"?

b) qual è l'ambito spazio-temporale di disponibilità della serie storica "y"?

c) esiste il dato "z"?

Questi quattro criteri di differenziazione, combinati fra loro, danno luogo a dieci tipi di richieste alla formulazione di ciascuna delle quali l'utente può arrivare più o meno direttamente a seconda del carattere più o meno preciso delle sue informazioni e delle sue esigenze iniziali

I cammini possibili attraverso cui un utente può esplorare l'Archivio sono rappresentati nello schema seguente:



Come si vede il vincolo principale del sistema è costituito dall'obbligo di precisare di quale data base di dati -D1 o D2- interessa conoscere il contenuto. Esso impone all'utente che non sappia su quale data base cercare, o che desideri cercare su entrambi, di effettuare due ricerche separate.

Questo vincolo naturalmente potrebbe anche essere rimosso disegnando in modo diverso l'ambiente di ricerca. Tuttavia un'eventualità del genere appare sconsigliabile per almeno due motivi:

a) i due tipi di data base di dati interessano presumibilmente due fasce diverse di utenti, o quanto meno due fasi distinte del lavoro del medesimo utente
b) una ricerca simultanea su due data base di contenuto così diverso renderebbe l'interazione con l'utente decisamente più complicata e faticosa, soprattutto negli stadi intermedi e finali.

Un'ultima osservazione sull'accessibilità reciproca dei vari stadi di una ricerca. Come si vede dallo schema di pag. .. alcuni rami dell'albero che rappresenta le selezioni successive effettuate nel corso di una ricerca sono intercomunicanti. Ciò significa che l'utente può saltare da un tipo di richiesta ad un altro senza tornare alla radice dell'albero. Questa possibilità è importante soprattutto nei casi in cui le esigenze e le aspettative dell'utente si aggiustano o si formano sulla base delle risorse e delle informazioni del sistema: l'utente capisce quel che vuole informandosi su quel che c'è. Possiamo esemplificare questo processo con una sequenza-tipo:

S: Su che data base vuoi lavorare?
U: Sul data base dei dati originari
S: Qual'è l'ambito spazio-temporale della tua richiesta?
U: Gli ultimi venti anni in Piemonte
S: Vuoi sapere che cosa c'è nel data base o hai una richiesta specifica?
U: Voglio sapere che cosa c'è
S: Le aree tematiche sono queste ... (lista) Quale ti interessa?
U: Mi interessa l'area tematica "x"
S: Gli indicatori sono questi ... (lista) Te ne interessa uno in particolare?

U: Mi interessa l'indicatore "y"
S: L'indicatore "y" ha le seguenti varianti ... (lista) Nell'ambito spazio-temporale da te definito la disponibilità di dati è la seguente ... (tabella)

A questo punto l'utente sa che cosa può trovare nel sistema fino al livello del dato singolo. Ciò non significa -beninteso- che l'utente abbia accesso al dato singolo dall'ambiente di Informazione. L'utente sa che in D1 c'è il dato che gli interessa, non qual è il valore numerico del dato stesso. Per conoscere quest'ultimo deve uscire dall'ambiente di Informazione ed entrare nell'ambiente "logicamente successivo", l'ambiente di Analisi dei dati.

2.5.3. Analisi dei dati

In ambiente di analisi dei dati l'utente può accedere direttamente ad un sottoinsieme dei dati stessi secondo una serie di percorsi più o meno rigidi. L'accesso ai dati è mediato da una serie di procedure altamente parametrizzate che operano direttamente sui data base D1 e D2 e abilitano l'utente a compiere tre famiglie di operazioni sui dati:

A. Analisi statistiche e grafiche elementari
B. Generazione di nuove serie storiche
C. Predisposizione di data set e system file di lavoro

Per analisi statistiche e grafiche elementari si intendono tutte le operazioni standard che si possono effettuare su una singola variabile sia in cross-section sia in time series:

- visualizzazione del contenuto del vettore, o del dato singolo
- analisi statistiche univariate (UNIVARIATE in SAS)
- plot contro il tempo
- analisi spettrale, filtraggio, estrapolazioni, simulazioni e simili.

Per generazione di nuove serie storiche si intendono due

differenti gruppi di operazioni.

Il primo è costituito dall'insieme di procedure di omogeneizzazione, unificazione, raccordo fra serie storiche che utilizzano esclusivamente successioni ordinate di operazioni elementari. In concreto ciò significa che l'utente che intende sottoporre una o più serie storiche a manipolazioni elementari può, se lo desidera, esimersi dallo scrivere e dall'eseguire i relativi programmi e limitarsi a fornire al sistema i parametri del suo problema attraverso una rapida interazione guidata da pochi menù.

Il secondo gruppo di operazioni è costituito dall'insieme di operazioni che consentono di standardizzare indicatori singoli, o di costruire indicatori composti o misti a partire da due o più indicatori.

La predisposizione di data set o system file di lavoro, infine, è anch'essa il risultato di una rapida interazione sistema-utente guidata da menù. In essa l'utente si limita a specificare le caratteristiche fondamentali del data set su cui intende lavorare:

- unità di analisi, che può essere sia di tipo spaziale sia di tipo temporale
- lista delle variabili, che possono provenire sia da D1 sia da D2.

In risposta il sistema gli mette a disposizione un system file o un data set già pronto, con tutti i nomi e le etichette delle variabili predisposte. Da questo punto in poi l'utente ha "sotto i piedi" una matrice case by variable su cui lavorare e davanti a se tutta la strumentazione -librerie di procedure, comandi e programmi- caratteristica del package nel cui linguaggio il data set è stato predisposto.

3. REALIZZAZIONE INFORMATICA

3.1. Premessa

Finora abbiamo parlato dell'Archivio degli indicatori sociali essenzialmente da un punto di vista esterno, cercando di definire che cosa l'Archivio deve essere in grado di fare. In questa seconda parte del progetto cercheremo di definire invece soprattutto come l'Archivio deve essere organizzato internamente per poter svolgere le funzioni che gli abbiamo assegnato. Questo non significa, tuttavia, che finora si sia parlato degli aspetti logici e che d'ora in poi si parlerà degli aspetti fisici, o che finora si sia parlato delle funzioni e che d'ora in poi si parlerà di scelte hardware e software. Il confine tra logico e fisico, fra architettura funzionale e scelte informatiche è sempre piuttosto difficile da tracciare, e lo è tanto più in una situazione di interazione strettissima fra evoluzione del software ed evoluzione dell'hardware quale quella che stiamo vivendo negli ultimi anni. Si pensi, solo per citare due esempi estremi, all'influenza di PROLOG sull'hardware dei calcolatori di quinta generazione, o all'incentivo che le prestazioni degli ultimi microprocessori per personal computer stanno fornendo alla "riconversione" dei grandi packages statistici (SAS e SPSS sotto MS-DOS).

Per questo ordine di considerazioni in quel che segue parleremo sì di scelte informatiche, ma ad un livello relativamente astratto. Quando ci riferiremo a specifiche scelte hardware -OH, personal IBM compatibili, ecc. -o a specifiche scelte software - SAS, DB3, Basic, Lisp- lo faremo a titolo puramente esemplificativo, per "rendere l'idea" del tipo di sistema che si intende realizzare e non per indicare opzioni tecniche vincolanti. L'unica scelta informatica su cui ci sentiamo finora di indicare un'opzione difficilmente reversibile è quella del package SAS come ambiente privilegiato per la gestione dei data base di dati (D1 e D2) e in genere per tutte le operazioni di tipo statistico (servizi all'utente). Il package SAS

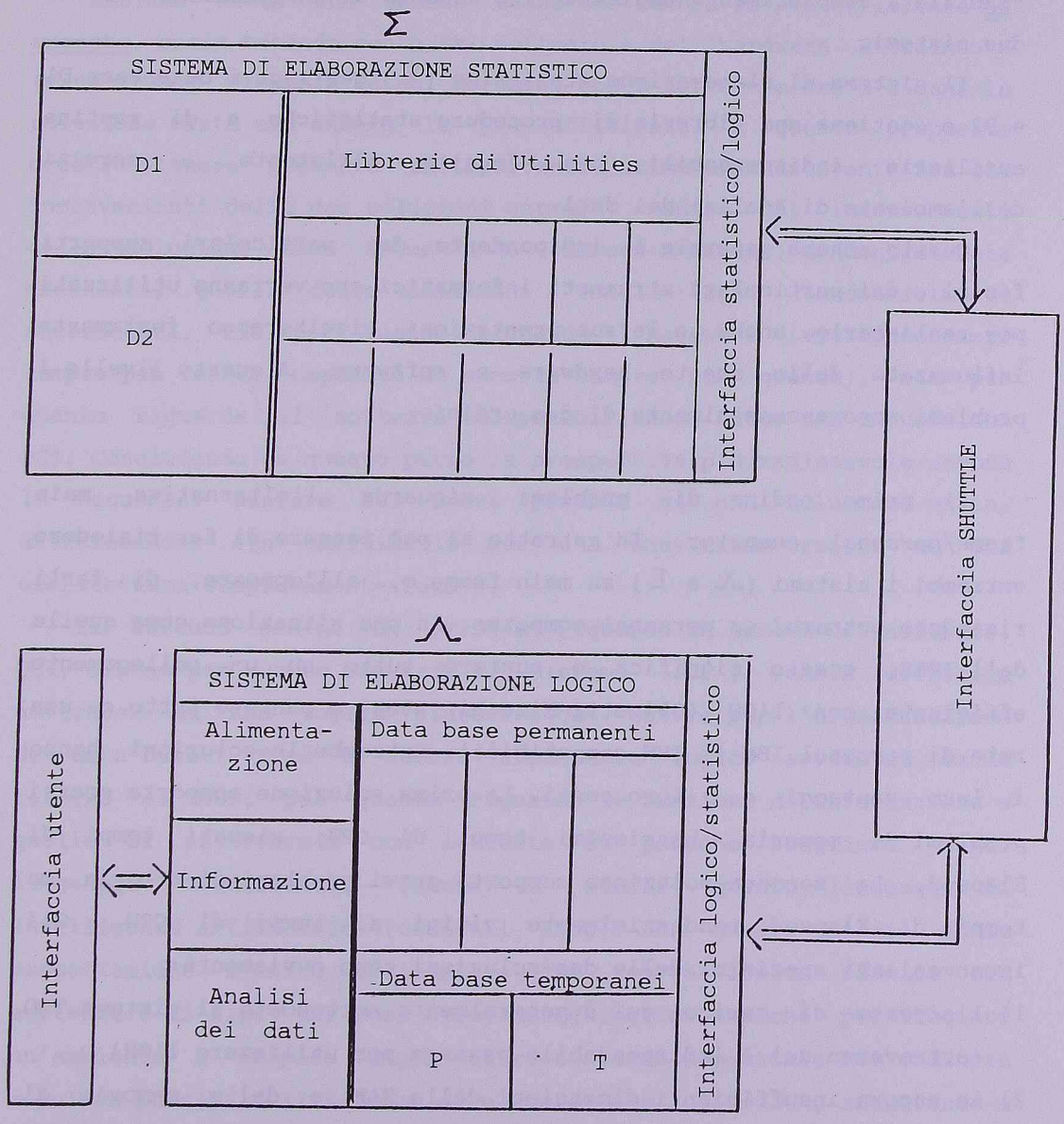
possiede infatti due requisiti che sono assai preziosi nel contesto di un progetto di archivio come quello delineato, e non sono rintracciabili in altri packages:

a) elevate capacità di programmazione

b) incorporazione del linguaggio APL.

3.2. Due sistemi di elaborazione e due interfaccia

Il funzionamento globale dell'Archivio può essere rappresentato attraverso lo schema seguente:



Le componenti base dello schema sono i due sistemi di elaborazione, uno logico (Λ) ed uno statistico (Σ).

Il sistema di elaborazione logico (Λ) gestisce i data base I, V ed S. Esso comunica con l'utente mediante l'interfaccia-utente, e con il sistema di elaborazione statistico mediante l'interfaccia "Shuttle", completamente dedicata allo scambio di informazioni fra i due sistemi.

Il sistema di elaborazione statistico (Σ) gestisce i data base D1 e D2 e contiene una libreria di procedure statistiche e di routine ausiliarie indispensabili per fornire all'utente i servizi dell'ambiente di Analisi dei dati.

Questo schema generale è indipendente dai particolari supporti fisici e dai particolari strumenti informatici che verranno utilizzati per realizzarlo, anche se le sue prestazioni risulteranno fortemente influenzate dalle scelte hardware e software. A questo livello i problemi sono essenzialmente di due ordini.

Il primo ordine di problemi riguarda l'alternativa main frame/personal computer. In astratto si può pensare di far risiedere entrambi i sistemi (Λ e Σ) su main frame o, all'opposto, di farli risiedere entrambi su personal computer. In una situazione come quella dell'IRES, questo significa o puntare tutto su un collegamento efficiente con l'OH (Olivetti-Hitachi 5560) o puntare tutto su una rete di personal IBM (o IBM compatibili). Entrambe le soluzioni hanno i loro vantaggi e i loro costi. La prima soluzione comporta scarsi problemi di memoria, bassissimi tempi di CPU, elevati tempi di Elapsed. La seconda soluzione comporta gravi problemi di memoria, ma tempi di Elapsed tendenzialmente vicini ai tempi di CPU. Gli inconvenienti specifici delle due soluzioni sono ovviamente:

1) l'eccesso di carico cui è generalmente sottoposto il sistema TSO (attraverso cui è indispensabile passare per utilizzare l'OH)
2) le ancora insufficienti dimensioni della RAM e delle memorie di massa dei personal computer.

In realtà la soluzione più efficiente, al momento, sembra una soluzione mista, che prevede una "divisione del lavoro" fra risorse

locali e risorse remote. Il sistema di elaborazione statistico (Σ) potrebbe risiedere sull'OH, ed essere gestito da SAS. Il sistema di elaborazione logico (Λ) potrebbe risiedere su personal computer, sotto MS-DOS, ed essere costituito da una serie di moduli eseguibili generati mediante programmi scritti in Basic, DB3 e Lisp.

Abbiamo detto che questa soluzione sembra la più efficiente al momento, ossia tenendo conto del software e dell'hardware attualmente (fine 1986) disponibile sul mercato. Non è detto che fra 2-3 anni la soluzione mista sia ancora la più efficiente. A questo proposito occorre tenere presente un'importante asimmetria esistente fra gli inconvenienti delle due soluzioni estreme: mentre gli inconvenienti della soluzione main frame sono (almeno a Torino) tendenzialmente crescenti, quelli delle soluzioni locali sono tendenzialmente decrescenti, sia per quanto riguarda l'hardware (microprocessori sempre più veloci, capacità di memoria sempre più grandi), sia per quanto riguarda il software (disponibilità di tutto SAS in versione PC). Concludendo su questo punto la prospettiva più ragionevole sembra la seguente: puntare nel breve periodo su una soluzione mista, attrezzandosi fin dall'inizio per una sua riconversione ad una architettura completamente locale.

Il secondo ordine di problemi riguarda la scelta del software. Qui, almeno per quanto riguarda il sistema Λ, sembra difficile sottrarsi ad una logica "pluralistica". Mentre per l'organizzazione dei data base I, V ed S sembra abbastanza logico avvalersi delle risorse di DB3, per quanto riguarda le funzioni di Informazione e quelle di interfaccia con l'utente si possono scegliere tanto linguaggi tradizionali (Basic e simili) quanto linguaggi di intelligenza artificiale (dialetti Lisp). In questo stadio della progettazione l'unica indicazione che sembra lecito fornire sul software è di tipo, per così dire, metodologico. Anzichè privilegiare un'ambiente a scapito degli altri conviene progettare nel modo più modulare possibile il sistema di elaborazione logico Λ, sfruttando a fondo le risorse dei compilatori DB3, Basic e Lisp per rendere confrontabili soluzioni alternative in tutti i casi incerti.

3.3. I sistemi di elaborazione Λ e Σ

Daremo ora una breve descrizione del funzionamento dei due sistemi Λ e Σ .

3.3.1. Il sistema di elaborazione logico Λ

Le funzioni del sistema di elaborazione logico Λ sono sostanzialmente quelle descritte nel paragrafo della Parte I dedicato all'interazione con l'utente:

1. Alimentazione
2. Informazione
3. Analisi dei dati.

Viste dall'interno del sistema esse si presentano tuttavia in una prospettiva un po' differente, che richiede qualche specificazione ulteriore.

La funzione di alimentazione (inizializzazione e input) non funziona nel modo semplice e diretto con cui è stata descritta. Sia l'inizializzazione (di indicatori e varianti) sia l'input (di dati) non operano direttamente sui data base I, V, D1 e D2, ma si svolgono in due fasi distinte, l'una di tipo virtuale, o preparatoria, l'altra di tipo reale, o definitiva. Nella prima fase vengono aperti dei data base temporanei in cui vengono convogliate sia le descrizioni di indicatori e varianti sia i dati. Tali data base sono interni al sistema di elaborazione Λ .

In una seconda fase l'utente dà l'ordine di trasferire i dati nel sistema di elaborazione Σ . Se e solo se il trasferimento ha successo il sistema compie le tre operazioni fondamentali di chiusura del ciclo di alimentazione:

a) ordine di scratch sui data base di <u>dati</u> residenti in Λ
b) aggiunta ad I e V delle descrizioni delle serie storiche appena inviate in Σ
c) ordine di scratch sui data base di <u>informazioni</u> residenti in Λ .

La funzione di Informazione si avvale esclusivamente del contenuto dei data base definitivi I, V e S, ed opera quindi in modo del tutto indipendente da Σ . Per sapere che cosa c'è in D1 e D2 non occorre far "viaggiare" l'interfaccia Shuttle, ma basta sfruttare i risultati dei suoi viaggi passati.

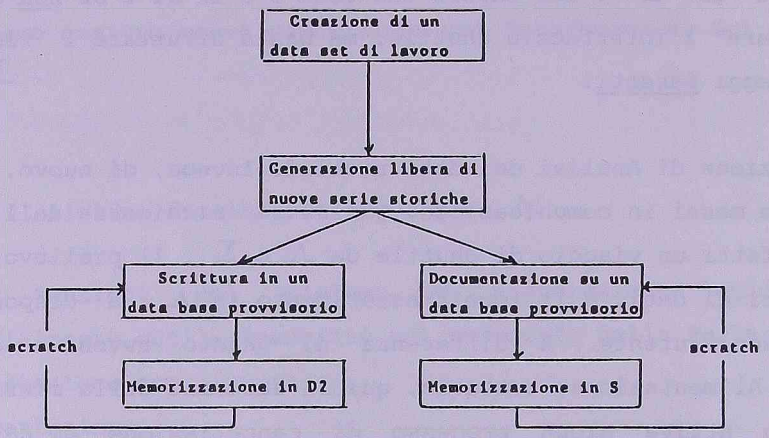
La funzione di Analisi dei dati richiede invece, di nuovo, che Λ e .. vengano messi in comunicazione. Qualsiasi richiesta dell'utente comporta infatti un viaggio di Shuttle da Λ a Σ , il prelievo di uno o più vettori di dati, e il loro trasferimento in Λ , a disposizione dell'interfaccia-utente. A differenza di quanto avveniva con la funzione di Alimentazione, tuttavia, qui il successo della missione di Shuttle non attiva alcun processo di cancellazione di data base provvisori. L'informazione in ritorno da Σ è infatti, a seconda dei casi, dei tre seguenti tipi:

a) display di dati, grafici, elaborazioni statistiche residenti in RAM
b) scrittura di serie storiche derivate in uno speciale data base di parcheggio P da cui possono essere usate oppure memorizzate definitivamente (scritte in D2 e documentate in S)
c) scrittura su video o su floppy di uno specifico programma di ingresso in SAS (data set di lavoro + ingresso in ambiente SAS).

In ciascuno dei tre casi precedenti le aree di memoria attivate al termine di un viaggio di ritorno da Σ a Λ vengono cancellate automaticamente con la conclusione della seduta di lavoro.

Nel disegno complessivo dei vincoli dell'Archivio resta aperto un problema che lasciamo volutamente in sospeso: possono serie storiche generate lavorando liberamente su data set SAS essere convogliate nel data base D2 delle serie storiche derivate, documentato in S?

Se si risponde affermativamente si è costretti a rendere notevolmente più complicata la struttura dell'ambiente di Analisi dei dati, che deve poter gestire sequenze del tipo:



Se si risponde negativamente alla domanda, e si vieta quindi all'utente di trasferire nuove serie storiche direttamente da un data set SAS liberamente costruito al data base D2 occorre rendere molto sofisticata la funzione Generazione di nuove serie storiche interna all'ambiente di Analisi dei dati, e metterla direttamente in comunicazione con l'ambiente di Alimentazione, consentendo sequenze del tipo:

1. Ingresso in ambiente di Analisi dei dati.
2. Selezione dell'opzione: Generazione di nuove serie storiche.
3. Attivazione dell'opzione e produzione effettiva di nuove serie storiche, che vengono scritte in P e documentate in un data base temporaneo T creato automaticamente.
4. Ingresso in ambiente di alimentazione e memorizzazione definitiva delle nuove serie storiche e della relativa documentazione in D2 e S.

3.3.2. Il sistema di elaborazione statistico Σ

Il sistema di elaborazione statistico contiene tre famiglie di "oggetti":

1. Dati
2. Procedure utente altamente parametrizzate
3. Procedure di sistema per la memorizzazione definitiva dei dati.

L'ambiente al cui interno ciascuno dei tre oggetti precedenti risiede è SAS. Ciò significa che i dati sono organizzati come archivi SAS, mentre le procedure sono per lo più macro SAS i cui parametri vengono istanziati dalle richieste che l'utente formula nell'ambito del sistema di elaborazione logico Λ .

Gli archivi SAS sono costituiti da un certo numero di data set SAS organizzati secondo il classico schema case by variable. E' importante osservare che la struttura logica delle matrici dati finora descritte non coincide con la loro organizzazione fisica. I data base D1 e D2 sono stati caratterizzati come insiemi di matrici bi e tridimensionali. I data set SAS che conterranno i dati rappresentati come D1 e D2 avranno una struttura alquanto differente da questi ultimi. Poichè tale struttura risulterà trasparente all'utente finora abbiamo descritto Σ come costituito dai data base D1 e D2. In questo paragrafo daremo invece una descrizione di Σ dal punto di vista della sua organizzazione effettiva, ossia come insieme di data set SAS. Nei data base D1 e D2 ogni matrice dati corrispondeva ad una serie storica (variante di indicatore o serie derivata). Inoltre tutte le matrici dati di D1 erano tridimensionali (tempo x spazio x livello di affidabilità).

Nei data set SAS, invece, tutte le matrici sono bidimensionali e seguono l'organizzazione standard case by variable. Esse sono costruite secondo queste regole:

Regola 1 Il record può essere soltanto un'unità temporale (mese, anno, ecc.) o spaziale (comune, provincia, regione).

Regola 2 Il record è un'unità temporale quando:

- la serie storica ha una cadenza regolare e al più annuale
- la serie storica viene registrata in meno di 10 unità spaziali.

In tutti gli altri casi il record è un'unità spaziale.

Regola 3 Conseguentemente ogni variabile contenuta in un generico data set SAS può essere caratterizzata univocamente mediante le seguenti caratteristiche:

1. Area tematica della serie storica corrispondente.
2. Unità di analisi (mese, anno, comune, ecc.), o significato del record nel data set SAS.
3. Carattere originario o derivato della serie storica corrispondente.
   3.1. Livello di affidabilità dei dati che costituiscono la serie (se la serie è originaria).
4. Nmero d'ordine dell'indicatore o della serie derivata nei data base I o S (le serie sono codificate in modo misto: sigla dell'area tematica più numero d'ordine nell'ambito del'area; es.: B007).
   4.1. Numero d'ordine della variante (se la serie è originaria) nel data base V.
5. Riferimento spaziale (se il data set SAS è time series) o temporale (se il data set SAS è cross-section).

Si potrebbe pensare di concentrare tutta l'informazione che caratterizza una variabile nel nome SAS della variabile stessa. In realtà, a parte le difficoltà tecniche insite in questo procedimento (non è detto che l'insieme delle variabili SAS possibili abbia molteplicità comprimibile nel contenuto d'informazione di una stringa di 8 caratteri alfanumerici), sembra preferibile ricostruire il contenuto di informazione di una variabile SAS utilizzando due fonti di informazione indipendenti:

a) il nome della variabile

b) il nome del data set SAS in cui la variabile è registrata.

Conviene, in altre parole, "distribuire" il contenuto d'informazione 1-5 in due famiglie di regole di formazione dei nomi: le regole per stabilire i nomi dei data set SAS, e le regole per stabilire i nomi delle variabili SAS interne a ciascun data set. Questo procedimento non solo consente di semplificare drasticamente la strutura dei nomi delle variabili SAS ma, se attuato seguendo determinati principi di economia, consente anche di abbreviare enormemente i tempi di attesa dell'utente.

Ma quali sono i principi di economia che possono guidare lo splitting del contenuto d'informazione 1-5?

Essenzialmente essi si riducono al principio guida seguente: rendere minimo il numero di data set SAS cui occorre accedere per soddisfare la medesima richiesta dell'utente. Tentativamente, possiamo proporre questa traduzione del principio precedente.

1. Fra le richieste che comportano un interazione Λ - Σ le più frequenti dovrebbero, in linea di mssima, rientrare nelle categorie seguenti:

A. Memorizzazione definitiva di blocchi di indicatori dotati della medesima organizzazione spaziale o temporale.
B. Ricucitura - raccordo - pulizia di varianti del medesimo indicatore.
C. Generazione di serie storiche derivate.
D. Preparazione di data set di lavoro dotati della medesima organizzazione spaziale o temporale.

2. Di norma i 4 tipi di richiesta A-D dovrebbero coinvolgere variabili SAS omogenee quanto a certi tipi di caratteristiche ed eterogenee rispetto ad altre. Le richieste di tipo A, ad esempio (memorizzazione definitiva), dovrebbero coinvolgere variabili SAS omogenee rispetto a 1, 2, 3, 4.1. ma non rispetto a 4. Le richieste di tipo B (ricucitura) dovrebbero coinvolgere variabili SAS omogenee rispetto a 1, 3, 4, 5 ma non rispetto a 3.1. E così via.

3. Se tentiamo di costruirci un'idea globale delle relazioni fra tipi di richieste e caratteristiche delle variabili SAS normalmente coinvolte in ciascuna richiesta, perveniamo ad uno schema del tipo seguente:

	1	2	3	3.1	4	4.1	5
A	=	=	=	?	≠	=	?
B	=	?	=	≠	=	≠	
C	?	=	=	?	≠	?	?
D	?	?	=	=	≠	≠	?
TOT "="	2	2	4	1	1	1	

Come si vede solo le caratteristiche 3, 2 e 1 (carattere originario/derivato, unità di analisi, area tematica) sembrano tendenzialmente condivise dalle variabili SAS che possono entrare in una richiesta dell'utente. Questo, in termini più concreti, vuol dire: quando un utente si siede a terminale e formula una richiesta (A, B, C, D) è più probabile che le variabili SAS che la soddisfano abbiano in comune le caratteristiche 3, 2 o 1 piuttosto che le atlre (3.1, 4, 4.1, 5).

4. Di qui una possibile indicazione strategica: collochiamo nel medesimo data set SAS tutte le variabili che hanno in comune l'area tematica (caratteristica 1), l'unità di analisi (caratteristica 2), e il carattere originario o derivato (caratteristica 3).

Questo criterio dà luogo ad un insieme di 10x2x(8+4)=240 data set SAS logicamente possibili, che si possono caratterizzare in modo comodo ed intuitivo mediante nomi di quattro caratteri.

Area tematica	Serie orig./ deriv.	Data set time series/ cross-section	Unità di analisi
A =	1 = origin.	T = time series	1 = mese
B =	2 = deriv.		2 = bimestre
...			.....
L =			8 = anno
		S = cross-section	1 = comune
			2 = provincia
			3 = regione
			4 = Italia/estero

Esempio. Il data set SAS A1T8 conterrà tutte le variabili SAS che:

- si riferiscono all'area tematica "Comportamenti e strategie famigliari" (A)
- sono costituite da serie originarie (1)
- sono organizzate in time series, con cadenza regolare (T)
- hanno cadenza annuale (8).

5. Dentro ogni singolo data set SAS "convivono" variabili che possono differire per le caratteristiche 3.1., 4, 4.1, 5. Non è detto che il nome SAS di ogni variabile debba contenere tutta l'informazione associata a tali caratteristiche. Si può anche dare ad ogni variabile SAS un nome sequenziale (V1, V2, ..., VN) e ricostruirne il contenuto di informazione mediante una tabella nomi --> vettori di

caratteristiche. Se viceversa si ritiene che sia utile concentrare l'informazione relativa ad una variabile SAS nel suo nome, allora occorrerà studiare un sistema di codifica che permetta di rappresentare in 8 caratteri le informazioni seguenti:

- affidabilità (3.1)
- numero d'ordine della serie (4)
- numero d'ordine della variante (4.1)
- riferimento spaziale o temporale (5).

Poichè alcune caratteristiche (3.1 e 4.1) sono condizionate, hanno cioè significato solo se le serie cui si riferiscono sono originarie, mentre altre (4 e 5) cambiano di significato a seconda del tipo di serie storica cui si riferiscono, converrà considerare separatamente i quattro tipi fondamentali di data set SAS (time series/cross-section, serie originarie/derivate) e definire un insieme di regole di codifica per ogni tipo di data set.

Vediamo ora brevemente l'altro tipo di files che costituiscono il sistema di elaborazione statistico Σ : procedure utente e procedure di sistema.

Le prime possono essere viste come membri di una libreria che permette al sistema di elaborazione logico Λ di rispondere a domande poste in ambiente di Analisi dei dati. E' il caso di ricordare che tutte le funzioni interne all'ambiente di Analisi dei dati possono essere svolte soltanto accedendo ai data set SAS che contengono i dati stessi. L'interazione sistema-utente segue, di norma, lo schema seguente:

1. Ingresso in ambiente di Analisi dei dati (utente).
2. Scelta di un'opzione (utente).
3. Passaggio dei parametri dell'opzione (utente).
4. Sintesi automatica del programma SAS che predispone il data set appropriato (sistema).
5. Sintesi automatica del programma SAS che esegue le istruzioni specifiche fornite al passo 3 (sistema).

6. Concatenazione in un programma SAS unico dei programmi sintetizzati ai punti 4 e 5 (sistema).
7. Invio del programma completo da Λ a Σ mediante l'interfaccia Shuttle (sistema).
8. Esecuzione del programma in Σ (sistema).
9. Trasmissione dell'output da Σ a Λ mediante Shuttle (sistema).
10. Visualizzazione o stampa dell'output su devices fisici locali: video, dischetto o carta (sistema).

Le procedure di sistema operano in modo analogo alle procedure utente, ma sono in tutto o in parte trasparenti (invisibili) all'utente stesso. Esse vengono attivate allorchè dall'ambiente di Alimentazione viene emesso un'ordine di memorizzazione definitiva di serie storiche, originarie o derivate, scritte e documentate in data base temporanei interni al sistema di elaborazione logico Λ . In tali circostanze si attiva una sequenza del tipo:

1. Ingresso in ambiente di Alimentazione (utente).
2. Selezione dell'opzione di memorizzazione definitiva (utente).
3. Dichiarazione del nome della matrice dati (o tabella) che contiene i dati da memorizzare (utente).
4. Check automatico sulla completezza e consistenza della documentazione relativa (sistema).
5. Generazione della lista di data set SAS che è necessario aprire o cui è necessario accedere per memorizzare i dati (sistema).
6. Sintesi automatica di un programma SAS che distribuisce i dati nei data set "giusti", e li scrive nelle righe e nelle colonne "giuste" (sistema).
7. Invio della matrice dati da leggere e del programma SAS di lettura-scrittura da Λ a Σ mediante Shuttle (sistema).
8. Esecuzione del programma in Σ (sistema).
9. Trasmissione da Σ a Λ di messaggi sull'esito delle operazioni di scrittura nei data set SAS di Σ .
10. Cancellazione (in caso di esito positivo) dei data base temporanei di dati (sistema).
11. Cancellazione (in caso di esito positivo) dei data base temporanei

di informazioni, e trasferimento del loro contenuto nei data base definitivi I e S (sistema).

12. Invio all'utente (in caso di esito negativo) di messaggi di errore.

Un'ultima osservazione, per concludere la descrizione di Σ, sulle operazioni di traduzione sistema --> utente e utente --> sistema che coinvolgono matrici di dati. Esse sono rese indispensabili da due circostanze:

a) in fase di Alimentazione l'organizzazione dei dati disponibili può essere radicalmente differente dalla loro organizzazione fisica finale come data set SAS
b) in fase di Analisi dei dati il tipo di data set su cui l'utente desidera lavorare può possedere una struttura radicalmente diversa da quella dei data set SAS da cui i dati provengono.

Più esattamente, le situazioni che si possono presentare sono le due seguenti:

a) un utente prepara una matrice d'ingresso in cui il record è di tipo spaziale, ma le serie storiche coinvolte devono essere memorizzate in data set SAS di tipo time series (o viceversa: record temporale, e serie storiche da memorizzare in cross-section)
b) un utente desidera lavorare in cross-section su un gruppo di variabili memorizzate in data set SAS di tipo time series (o viceversa).

Queste eventualità suggeriscono due requisiti che i sistemi di elaborazione Λ e Σ devono possedere.

In primo luogo il sistema di elaborazione statistico Σ deve essere in grado di gestire il cambiamento di unità di analisi mediante procedure di sistema -per lo più trasparenti all'utente- che convertono uno o più data set con una data unità di analisi in data set con un'unità di analisi differente.

In secondo luogo il sistema di elaborazione logico Λ deve essere in grado di eseguire automaticamente sulle matrici dati di input tutte

quelle operazioni di trasposizione e riordino di linee che permettono di leggerle in un formato prossimo a quello finale (data set SAS definitivi). E' questa la funzione principale delle regole di ingresso dei dati descritte nell'Appendice 5.

3.4. Le interfaccia

I due sistemi di elaborazione Λ e Σ costituiscono il nucleo dell'Archivio, nel senso che in essi è concentrata tutta la competenza necessaria per svolgere le varie funzioni finora descritte. E tuttavia il nostro sistema non deve essere semplicemente in grado di assolvere determinati compiti. Altrettanto importante è che tali compiti siano svolti con agilità, rapidità e affidabilità. Questi ultimi aspetti sono fortemente influenzati dalla struttura delle due interfaccia fondamentali dell'Archivio: l'interfaccia utente e l'interfaccia Shuttle.

3.4.1. L'interfaccia utente

Quando si progetta un'interfaccia utente occorre, di norma, effettuare una scelta tra due strategie fondamentalmente diverse:

a) dialogo in linguaggio naturale
b) dialogo guidato da menù

Nel nostro caso la scelta è più apparente che reale, perchè la prima alternativa è praticamente irrealizzabile. I sistemi capaci di dialogare con l'utente accettando richieste in linguaggio naturale presentano infatti un insieme di limitazioni che vanificherebbero il progetto stesso di un Archivio degli indicatori sociali dotato delle capacità finora descritte. Fra queste limitazioni è il caso di ricordare:

a) competenza linguistica su domini ristretti
b) difficoltà di gestire il cambiamento di ambiente, ovvero di cogliere il "contesto" delle richieste
c) allungamento dei tempi di attesa dovuto alla pesantezza e alla complessità delle operazioni di parsing, interpretazione ecc.
d) elevati tempi di progettazione e sperimentazione.

Se la scelta di un tipo di interazione guidata da menù è

praticamente una scelta obbligata, meno ovvia appare la scelta dei modi concreti di realizzarla. A questo livello si possono già indicare alcune caratteristiche che l'interfaccia utente dell'Archivio dovrebbe possedere. La prima è l'adozione di un formato il più uniforme possibile per i pannelli, e di pochissimi standard di interazione:

- selezione di un'opzione mediante digitazione di un numero di una cifra
- scorrimento di pannelli di informazioni ausiliarie
- richiesta di aiuto (help)
- percorrimento a ritroso della sequenza dei menù.

La seconda caratteristica è la valorizzazione e l'evidenziazione della nozione di ambiente. Questo significa da un lato che l'utente deve sempre sapere in quale dei tre ambienti (nonchè degli eventuali sottoambienti) si trova, dall'altro che in determinati contesti di interazione devono esistere dei canali "laterali" di accesso diretto ad altri ambienti (l'utente deve, ad esempio, potere effettuare una ricerca nel contesto di una inizializzazione di indicatori).

L'ultima caratteristica, cui si è già accennato in un precedente paragrafo (par. 2, Parte II), è l'integrazione di strumenti di intelligenza artificiale (moduli eseguibili LISP) all'interno di un'architettura generale fondata sui menù. Si tratta, in altre parole, di rompere la semplicistica equazione intelligenza artificiale = linguaggio naturale, e di porre le capacità di manipolazione di stringhe proprie di linguaggi come LISP al servizio di un'interfaccia "tradizionale". Questo approccio appare particolarmente promettente in ambiente di Informazione, soprattutto in quei casi in cui le richieste dell'utente sono relativamente fuzzy e si prestano quindi ad essere precisate attraverso un dialogo basato su parole chiave.

3.4.2. L'interfaccia Shuttle

Il momento più delicato dell'intero funzionamento dell'Archivio è quello in cui l'utente dà l'ordine di memorizzazione defintiva di un

certo data base di dati residente (temporaneamente) in Λ . E' a questo punto che entra in gioco l'interfaccia Shuttle, che ha il compito di trasportare i dati da Λ a Σ , "trovare posto" in Σ per collocarli, nonchè tornare in Λ per aggiornare definitivamente i data base di informazioni (I, V o S) e comunicare all'utente l'esito delle operazioni.

Meno delicato ma altrettanto importante è il momento in cui l'utente formula una richiesta in ambiente di Analisi dei dati: poichè i dati risiedono in Σ e non in Λ , anche qui deve entrare in funzione l'nterfaccia Shuttle che trasmette la richiesta a .. e "trasporta" la risposta da Σ a Λ .

Il modo di realizzare effettivamente l'interfaccia Shuttle dipenderà in modo critico dal tipo di architettura (completamente remota, completamente locale, o mista) dell'Archivio. In questo stadio della progettazione è però già possibile delineare qual è la funzione essenziale dell'interfaccia, indipendentemente dalle scelte hardware e software finali. Tale funzione può essere enunciata così. L'interfaccia Shuttle ha il compito di assicurare contemporaneamente:

- la protezione dei data base definitivi residenti in Λ e Σ (I, V, S, D1, D2)
- la consistenza reciproca fra data base di informazioni (I, V, S) e data base di dati (D1, D2)
- la comunicazione (trasmissione di informazioni) fra i due sistemi di elaborazione logico e statistico (Λ e Σ).

Detto in termini più concreti: nessun utente deve essere in grado di aggiungere, modificare o aggiornare dei dati negli Archivi D1 e D2 senza essere passato attraverso le "forche caudine" della documentazione (inizializzazione di indicatori e varianti). D'altro canto, simmetricamente, nessun utente può descrivere su I, V o S -che sono data base di informazioni definitivi- qualcosa che non esiste ancora, o che esiste in modo difforme, in D1 o D2.

Questa triplice esigenza suggerisce di dotare i sistemi di elaborazione Λ e Σ di un potere di veto reciproco, che passa attraverso la conoscenza esclusiva delle password di accesso ai data

base definitivi: solo Λ sa come scrivere su Σ, e solo Σ sa se e come documentare in Λ ciò che è stato scritto su Σ.

Naturalmente questa capacità incrociata di veto richiede che i due sistemi Λ e Σ siano dotati di capacità crittografiche, e siano protetti contro l'ispezione dall'esterno di dette capacità. Si può, ad esempio, stabilire che le password di accesso dipendono dalla data e dall'ora secondo una o più regole di codificazione della data e dell'ora stesse, ma tali regole devono risultare del tutto inaccessibili agli utenti e, possibilmente, (almeno entro certi limiti) ai sistemisti stessi.

Appendice 1

Tipologia degli indicatori sociali

Abbiamo più volte richiamato la tipologia-base degli indicatori sociali:

A - Comportamenti e strategie familiari
B - Reati
C - Comportamenti elettorali
D - Consumi culturali
E - Scommesse
F - Comportamenti collettivi e associazionismo
G - Suicidio e altre cause di morte
H - Salute
I - Incidenti del traffico
L - Economia e mercato del lavoro.

E' ora il caso di precisare meglio la struttura generale e le articolazioni interne di questa tipologia.

In generale i vari tipi vanno considerati mutuamente esclusivi. Per evitare incertezze e arricchire la tipologia è quindi indispensabile prevedere, per ogni tipo, un certo numero di sottotipi più specifici, evitando di duplicare informazione mediante attribuzioni multiple. I suicidi e i morti per droga, ad esempio, vanno collocati nel tipo G (Suicidi e altre cause di morte) piuttosto che nel tipo H (Salute). Così gli scioperi vanno collocati nel tipo F (Comportamenti collettivi e associazionismo) piuttosto che nel tipo L (Economia e mercato del lavoro).

Queste attribuzioni sono convenzionali e vengono decise una volta per tutte a livello di sistema. L'utente che vuole inizializzare una nuova serie storica deve collocarla in modo univoco in un tipo e in un sottotipo. La struttura per tipi e sottotipi è fissata a priori mentre la lista di indicatori associata a ogni sottotipo è liberamente costruita dall'utente. Attualmente, a titolo puramente esemplificativo, si può proporre la seguente organizzazione per tipi e sottotipi:

A: Comportamenti e strategie familiari

- A1. Nati legittimi e illegittimi
- A2. Matrimoni religiosi e civili
- A3. Separazioni e divorzi
- A9. Altro

B. Reati

- B1. Omicidi e rapine
- B2. Reati contro il patrimonio
- B3. Reati contro la morale
- B4. Delitti contro leconomia pubblica, l'industria e il commercio
- B5. Delitti contro la fede pubblica
- B6. Delitti contro la personalità dello stato
- B9. Altro

C. Comportamenti elettorali

- C1. Astensioni
- C2. Schede nulle
- C3. Schede bianche
- C4. Voti a singoli partiti
- C9. Altro

D. Consumi culturali e istruzione

- D1. Istruzione
- D2. Partecipazione a spettacoli
- D3. Partecipazione a manifestazioni sportive
- D4. Letture
- D5. Ascolto radio e televisione
- D9. Altro

E. Scommesse

- E1. Totocalcio
- E2. Lotto
- E3. Totip
- E9. Altro

F. Comportamenti collettivi e associazionismo
- F1. Scioperi
- F2. Iscritti a sindacati
- F3. Iscritti a partiti
- F4. Iscritti a organizzazioni e movimenti religiosi
- F5. Iscritti ad altre organizzazioni
- F9. Altro

G. Suicidi e altre cause di morte
- G1. Suicidi
- G2. Tentati suicidi
- G3. Morti per droga
- G4. Morti sul lavoro
- G5. Morti in incidenti del traffico
- G6. Morti per cause accidentali

H. Salute
- H1. Malattie
- H2. Morti per malattie
- H3. Degenze ospedaliere

I. Incidenti del traffico
- I1. Incidenti
- I2. Persone infortunate
- I3. Veicoli coinvolti

L. Economia e mercato del lavoro
- L1. Reddito
- L2. Consumi
- L3. Investimenti
- L4. Risparmio
- L5. Prezzi
- L6. Salari e stipendi
- L7. Occupazione e disoccupazione

L8. Cassa integrazione

L9. Altro

Rispetto a questa tipologia l'utente può intervenire a tre livelli:

1. Aggiungendo serie storiche nuove all'interno di un sottotipo già previsto
2. Collocando nel sottotipo residuale "9. Altro" interno ad ogni tipo eventuali serie storiche nuove che rientrano nel tipo ma non sono previste da nessun sottotipo
3. Collocare in uno speciale tipo R (residuo) privo di struttura interna tutte le serie storiche che non solo non rientrano in nessun sottotipo esplicitamente previsto, ma neppure risultano collocabili in uno dei 10 tipi fondamentali (A-L).

Questa organizzazione generale è sufficiente per la inizializzazione di indicatori e varianti, ma non lo è per quanto riguarda la generazione e la documentazione automatica di serie storiche derivate. Qui può accadere che il tema di una serie storica sia, per così dire, strutturalmente misto o ambiguo, in quanto la serie è costruita "pescando" da indicatori collocati in aree tematiche differenti (esempio: omicidi/suicidi). In questo caso la serie andrà classificata in uno speciale tipo M (serie miste), i cui sottotipi potranno essere caratterizzati mediante stringhe di caratteri ottenute per combinazione delle 10 lettere base A-L (esempio: un indicatore costruito come rapporto suicidi/omicidi andrà collocato nel tottotipo BG, dove B rappresenta il tipo Reati, e G il tipo Suicidi e altre cause di morte).

Appendice 2

Descrittori degli indicatori

Per immettere dei dati nell'Archivio l'utente è tenuto a specificare -tra l'altro- quali sono gli indicatori cui i dati si riferiscono. Gli indicatori devono essere stati precedentemente inizializzati specificando il valore dei loro descrittori, che costituiscono i campi del data base I. I descrittori di un indicatore sono sei:

1. La sua collocazione nell'ambito della tipologia (tipo e sottotipo)
2. Il suo nome, che ne precisa il contenuto, a cui è associato un indice sequenziale
3. Un (eventuale) commento sull'indicatore stesso
4. La fonte da cui provengono i dati relativi all'indicatore
   Per fonte non si intende la fonte cartacea (estremi della pubblicazione) ma il nome o la sigla del soggetto istituzionale responsabile della prima pubblicazione dei dati
5. L'estensione della rilevazione (universo, campione, stima indiretta)
6. Il livello di scala della variabile associata all'indicatore

I livelli di scala ammessi sono tre:

I. Scale assolute (esempio: numero di addetti)
II. Scale di rapporti (esempio: reddito)
III. Scale di intervalli (esempio: temperatura)

Nel caso delle scale assolute occorre precisare anche qual è l'unità di conto. I casi più frequenti di unità di conto rientrano nei seguenti tipi:

- eventi (esempio: matrimoni)
- persone (esempio: numero di disoccupati)

- animali (esempio: capi di bestiame)
- cose (esempio: autoveicoli)

Nel caso delle scale di rapporti e di intervalli occorre precisare qual è l'unità di misura. Poichè, tuttavia, l'unità di misura è un possibile elemento di differenziazione fra varianti del medesimo indicatore (esempio: reddito in dollari e in lire) l'informazione ad essa relativa deve essere fornita durante il proceso di inizializzazione delle varianti (vedi Appendice 3).

Appendice 3

Descrittori delle varianti

Ogni indicatore possiede una o più varianti, che devono essere descritte accuratamente e documentate nel data base V. I descrittori delle varianti sono sette:

1. L'unità di misura (se la scala è di intervalli o di rapporti)
   Nell'unità di misura rientra anche l'eventuale espressione dei dati in decine, centinaia, migliaia, milioni, ecc. (caso molto frequente in contabilità nazionale)

2. La cadenza di rilevazione
   I tipi di cadenza previsti sono dieci, di cui nove regolari e una no:
   - giorno
   - settimana
   - mese
   - bimestre
   - trimestre
   - quadrimestre
   - semestre
   - anno
   - decennio
   - cadenza irregolare

3. Il tipo di misurazione (o di conteggio)
   Con questa espressione ci si riferisce alla distinzione fra rilevazioni puntuali (occupati nella settimana di riferimento), rilevazioni che sono medie di più rilevazioni puntuali (occupati in media annua: media delle rilevazioni di gennaio, aprile, luglio, ottobre), e rilevazioni che registrano gli integrali di determinati flussi entro un arco di tempo dato (chilowattora consumati nel corso di un bimestre). Abbiamo dunque tre casi:

- valori puntuali
- valori medi
- valori cumulativi

La distinzione è rilevante perchè comporta differenze di fondo nelle tecniche di aggregazione/disaggregazione temporale delle serie (esempio: per passare da dati mensili a dati annui si fa una media se la misurazione è puntuale, si fa una somma se la misurazione è cumulativa)

4. Periodo di riferimento

Qui i casi possibili sono dieci:

- giorno
- settimana
- mese
- bimestre
- trimestre
- quadrimestre
- semestre
- anno
- decennio
- altro

Il periodo di riferimento non va confuso con la cadenza. Il periodo di riferimento è l'intervallo o l'istante temporale al quale il dato si riferisce, la cadenza è la lunghezza dell'intervallo fra due rilevazioni consecutive. Esempi:

- i censimenti demografici si riferiscono allo stato della popolazione in un particolare giorno dell'anno ma hanno una cadenza decennale
- l'indagine sulle forze di lavoro si riferisce ad una particolare settimana di un dato mese ma ha una cadenza trimestrale

5. Definizione operativa della variante

Si tratta di una descrizione qualitativa dei tratti distintivi della definizione operativa della variante, eventualmente in contrapposizione ad altre varianti dotate di definizioni operative

differenti

6. Confini temporali della variante

Normalmente il periodo storico in cui "vige" una determinata variante può essere caratterizzato facilmente con un estremo iniziale (esempio: primo trimestre '59) e con un estremo finale (esempio: ultimo trimestre 1976), che può anche essere "aperto" (esempio: "fino a oggi, dicembre 1986").

Nei casi più complessi, in cui la rilevazione viene interrotta e poi ripresa, può essere necessario specificare più di un segmento temporale.

7. Estremi della pubblicazione

Si tratta degli estremi delle pubblicazioni ufficiali (volumi statistici) in cui sono comparsi il primo e l'ultimo dato di ogni segmento della serie.

Appendice 4

Operazioni sulle serie originarie

Il data base D2 contiene esclusivamente serie storiche ricavabili da serie storiche originarie contenute in D1 e documentate in A e V. Questo principio metodologico è fondamentale perchè garantisce la pulizia, la documentabilità e la trasparenza di tutto ciò che risiede nell'Archivio. Un generico "indicatore" di un certo fenomeno sociale non può legittimamente far parte dell'Archivio se non si è in grado di precisare in modo completo la sequenza di passi che ha consentito di costruirlo.

Questo principio impegna anche, tuttavia, a rendere estremamente ricco e sofisticato il modulo Generazione di nuove serie storiche, interno all'ambiente di Analisi dei dati. Ciò suggerisce di strutturare l'ambiente stesso come una collezione di operatori statistici combinabili e concatenabili in modo tale da consentire di generare automaticamente (direttamente sotto la guida dei menu) qualsiasi serie storica che possa essere derivata da serie storiche contenute in D1.

A titolo esemplificativo possiamo considerare i seguenti operatori fondamentali:

OP1. Caricamento di una o più serie originarie in un data set di lavoro
OP2. Cancellazione di segmenti temporali di una serie
OP3. Concatenamento fra serie per accostamento di segmenti temporali contigui
OP4. Raccordo fra segmenti eterogenei
OP5. Ritardo e avanzamento di serie (LAG)
OP6. Trasformazione di una serie in livelli in una serie in incrementi
OP7. Trasformazione di una serie in livelli in una serie in saggi di variazione (incrementi percentuali)
OP8. Cambiamento di unità di misura

OP9. Aggregazione temporale (esempio: dati trimestrali --> dati annui)

OP10. Disaggregazione temporale (esempio: dati annui --> dati trimestrali)

OP11. Destagionalizzazione

OP12. Depurazione del trend

OP13. Interpolazione

OP14. Estrapolazione

OP15. Trasformazioni matematiche ad 1 argomento (radice, elevamento a potenza, logaritmo naturale, ecc.)

OP16. Operazioni simmetriche su N argomenti

OP17. Operazioni asimmetriche su 2 argomenti (sottrazione, divisione, elevamento)

OP18. Media ponderata fra serie storiche

Naturalmente ogni operatore è anche caratterizzato da un vettore di parametri, che specificano i dettagli dell'operazione che deve compiere.

Questa impostazione comporta un importante vantaggio collaterale per l'architettura dell'Archivio. La documentazione di qualsiasi serie derivata può essere (quasi) interamente automatizzata. Anzichè attraverso una complessa struttura costituita da descrittori, la documentazione in S di una serie contenuta in D2 e generata a partire da D1 può essere ottenuta mediante due soli campi carattere, l'uno utile ma non strettamente indispensabile, l'altro obbligatorio ma generato automaticamente:

Campo 1. Descrizione sintetica del significato della serie e dei procedimenti mediante cui è stata costruita

Campo 2. Traccia simbolica (Argomenti + Operatori) delle operazioni, elementari e non, con cui la serie è stata generata.

Appendice 5

Formato di ingresso dei dati

Abbiamo visto nel paragrafo 5.1 che l'input di dati può avvenire secondo due modalità principali:

a) lettura da dischetto di matrici case by variable
b) digitazione diretta di tabelle residenti su carta

Vediamo ora più concretamente quali sono i vincoli semantici e sintattici che occorre rispettare, iniziando da quelli che valgono in entrambi i casi.

L'input di dati avviene secondo una sequenza fissa:

1) si definisce, dandole un nome, una matrice rettangolare di dimensione (relativamente) arbitraria
2) si specifica il significato delle righe e delle colonne e il livello di affidabilità dei dati
3) si scrivono i dati (facendoli leggere o digitandoli) nella matrice in questione, che viene salvata automaticamente come data base temporaneo in ⍀
4) quando la matrice è completa (il che in certi casi può avvenire anche in una seduta successiva a quella iniziale) si dà l'ordine di memorizzazione definitiva

E' il caso di precisare che la matrice può anche essere un vettore-riga (matrice 1 x n), un vettore-colonna (matrice n x 1) o uno scalare (matrice 1 x 1). Il punto essenziale è che l'immissione di dati rigaurda, in generale, insiemi di dati (relativamente) completi e denominati. Non si può cominciare a scrivere o far leggere dei dati se non si è prima definita una struttura a priori.

In generale è preferibile che i dati contenuti nella medesima matrice siano caratterizzati dal medesimo livello di affidabilità. Ciò abbrevia sia le operazioni di definizione della struttura di accoglimento dei dati, sia quelle di memorizzazione definitiva. Se il livello di affidabilità dei dati non è costante in tutta la matrice

l'utente ha a disposizione quattro possibili alternative:

a) suddividere la matrice originaria in due o più blocchi caratterizzati da livelli di affidabilità differenti e trattare ogni blocco come una matrice a sè stante, con un suo nome, una sua struttura a priori ecc.
b) invocare l'opzione che consente di definire il livello di affidabilità di ogni riga della matrice
c) invocare l'opzione che consente di definire il livello di affidabilità di ogni colonna della matrice
d) invocare l'opzione che consente di definire il livello di affidabilità di ogni cella della matrice

Un ultimo vincolo che coinvolge entrambe le modalità di ingresso dei dati riguarda il formato dei dati, che vengono sempre letti e scritti in doppia precisione, con al più tre cifre alla destra della virgola.

Consideriamo ora le differenze fra le due modalità di inputazione dei dati. Se si trascurano le differenze ovvie fra lettura da dischetto e digitazione diretta l'unica differenza importante è quella che riguarda la struttura della matrice di ingresso (fase 2 della sequenza tipo). Vediamo separatamente che cosa accade nei due casi.

Lettura da dischetto

L'utente deve, innanzitutto, dire che nome intende dare all'insieme di dati che sta per leggere. A questo punto il sistema gli chiede se il record (le "righe" della matrice da leggere) è di tipo spaziale o di tipo temporale. L'utente ha tre risposte a disposizione:

1. Record di tipo temporale
2. Record di tipo spaziale
3. Record di altro tipo

I primi due tipi di risposta sono i tipi normali. E' infatti piuttosto raro che dati di livello di scala alto (scale assolute, di rapporti, di intervalli) risultino registrati in matrici case by

variable in cui il record non rappresenta nè un'unità di tempo nè un'unità territoriale.

Dopo aver specificato la natura -temporale o spaziale- del record l'utente deve precisare:

- il numero di record
- il riferimento specifico (ex: gennaio '82, comune di Asti ecc.) di ogni record

A questo punto l'utente può iniziare a precisare la struttura del tracciato record o, più semplicemente, il significato di ogni colonna della matrice dati da leggere. La definizione del significato di ogni colonna avviene mediante un repertorio di codici preesistenti, che dipendono sia dal vocabolario interno del sistema sia dal lessico costruito dall'utente. Più esattamente:

- se il record è di tipo temporale l'utente deve fornire, per ogni colonna, il suo riferimento spaziale e i codici associati all'indicatore e alla variante corrispondenti
- se il record è di tipo spaziale l'utente deve fornire, per ogni colonna, il suo riferimento temporale e i codici associati all'indicatore e alla variante corrispondenti

Facciamo un esempio. Si tratta di informare il sistema, che già sa che il record è l'anno e il periodo di riferimento è il decennio 1971-1981, sul significato della colonna 3 del tracciato record. Supponiamo che la colonna in questione rappresenti i nati vivi totali in provincia di Novara (Indictore A021, Variante 1). Il dialogo sistema-utente si svolge pressapoco così:

S: Qual è il riferimento spaziale della colonna 3?

1. Un comune del Piemonte
2. Una provincia del Piemonte
3. Una regione italiana
4. L'Italia nel suo insieme
5. L'estero

U: 2

S: Quale provincia?

1. Alessandria
2. Asti
3. Cuneo
4. Novara
5. Torino
6. VercellA

U: 4

S: Qual è la sigla dell'indicatore rappresentato nella colonna 3?

U: A021

S: E il numero d'ordine della variante?

U: 1

Naturalmente, se tutte le colonne hanno il medesimo riferimento spaziale (temporale) oppure hanno il medesimo riferimento concettuale (indicatore e variante) l'utente può indicarlo al sistema e alleggerire così i termini del dialogo.

Il terzo tipo di risposta (3. Record di altro tipo) si può verificare quando:

a) i dati si riferiscono tutti al medesimo contesto spaziale e al medesimo momento di rilevazione (esempio: Piemonte, 1^ trimestre '83) e nel salto da un record all'altro cambia la natura di ciò che viene rilevato (cambia l'indicatore)
b) i dati non si riferiscono tutti al medesimo contesto spaziale e al medesimo momento di rilevazione ma tale variabilità differenzia fra loro le colonne anzichè le righe della matrice

In entrambi i casi precedenti (a e b) la matrice dati deve essere riorganizzata prima di essere letta. Nel caso a) il programma si farà dire il significato di ogni cella della matrice, e leggerà la matrice stessa come un record unico, a riferimento spazio-temporale costante. Nel caso b) il programma leggerà la trasposta della matrice anzichè la matrice originaria, e proseguirà il dialogo con l'utente in modo analogo a quanto accade dopo risposte di tipo 1 o 2.

A rigore, accanto ai casi a) e b), occorrerebbe considerare un terzo caso:

c) i dati non si riferiscono tutti al medesimo contesto spaziale e al medesimo momento di rilevazione ma nè ciò che differenzia fra loro le righe, nè ciò che differenzia fra loro le colonne può essere ridotto completamente a differenze di tempo o di spazio.

In questo terzo caso la matrice può essere letta soltanto specificando pazientemente, cella per cella, qual è il riferimento temporale, spaziale e concettuale (indicatore + variante) di ogni singolo dato.

Lettura da carta

Un buon programma di lettura di tabelle residenti su carta deve essere in grado di conciliare almeno tre esigenze in conflitto fra loro:

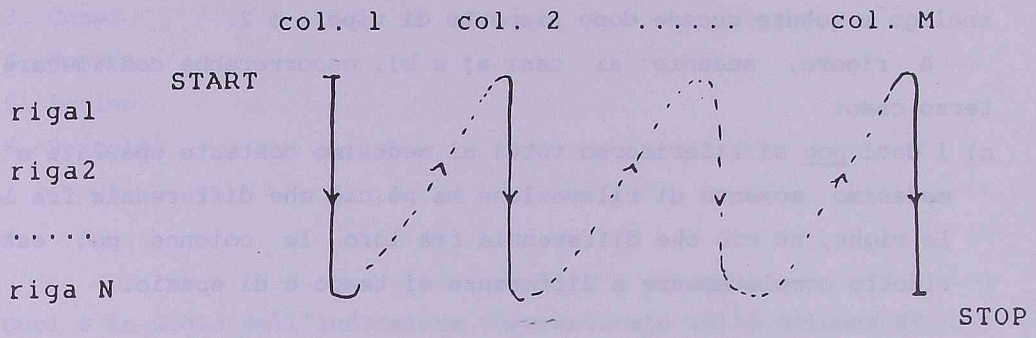
a) guidare l'immissione dei dati con una mascherina che riproduca il più fedelmente possibile la struttura della tabella
b) far stare tutta la tabella, o la frazione più grande possibile di essa, in una pagine a video (25 x 80 linee)
c) evitare tempi di attesa troppo lunghi fra due digitazioni successive

Senza pretendere di esaurire già in questa sede il problema ci limitiamo qui ad alcune indicazioni di massima.

1. Le regole di definizione di righe, colonne e celle singole dovrebbero essere il più possibile simili a quelle della lettura da

dischetto.

2. Poichè di norma la digitazione è più agevole e più rapida scendendo lungo le colonne che scorrendo orizzontalmente le righe conviene, salvo in casi molto particolari, far seguire al puntatore un tracciato di questo tipo:



3. La digitazione avviene una colonna per volta. La pagina video ha un titolo (significato della colonna) e -se la riga rappresenta unità temporali o spaziali- un'intestazione laterale. Se manca l'intestazione laterale il titolo della pagina rappresenta il contenuto del singolo dato digitato di volta in volta. Le informazioni necessarie per costruire il titolo sono ricavate automaticamente dai data base locali I e V. Il fatto che la digitazione avvenga una colonna per volta consente di registrare tabelle con un numero qualsiasi di colonne.
4. Quando una colonna ha più elementi del numero di righe disponibili sullo schermo il contenuto del video scorre automaticamente in linea verticale, consentendo la registrazione di tabelle di un numero qualsiasi di righe.

PARTE II UN APPROCCIO COSTRUTTIVISTA ALL' ORGANIZZAZIONE DEI DATI

PREMESSA

Lo scopo principale di questo secondo rapporto, che segue ad un anno di distanza il progetto di fattibilità, è di delineare in modo più preciso la teoria dei dati impliciti nell'architettura logica e funzionale dell'Archivio. La conoscenza e la familiarità con i concetti base con cui l'Archivio lavora è infatti essenziale per consentire all'utente un accesso alle informazioni efficiente ed "amichevole", nonchè un pieno utilizzo delle potenzialità dell'Archivio stesso.

L'esposizione è divisa in tre parti. Nella prima si presentano e si discutono brevemente i concetti base che permettono di percorrere la lunga catena che da una generica area tematica conduce fino al singolo dato. Nella seconda parte viene ripreso e sviluppato uno dei concetti chiave già introdotti nel progetto di fattibilità: il concetto di trasformazione elementare. Esso permette di dare una certa articolazione alla gerarchia di trasformazioni che dalle serie grezze conducono a livelli via via più complessi di strutturazione e riorganizzazione dei dati. Nella terza parte, infine, si affronta il problema dell'interazione sistema-utente cercando di mettere a fuoco le principali modalità secondo le quali esso si potrà svolgere. L'utilità dell'Archivio degli indicatori sociali non dipenderà infatti soltanto dalla quantità e dalla qualità dei dati in esso registrati ma anche, in modo cruciale, dalla capacità dell'utente di cogliere le possibilità del sistema e, simmetricamente, dalla capacità del sistema di interpretare correttamente le richieste dell'utente.

1. DAL TEMA AL DATO: UNA CATENA DI CONCETTI

Per passare dalla generica individuazione del tema di una determinata richiesta alla singola serie storica, alla singola cross-section, o addirittura al singolo dato, l'utente deve percorrere una catena di passaggi logici:

Schema generale

0. Area tematica
   1. Sottoarea tematica
      2. Indicatore sorgente
         3. Indicatore
            4. Variante
               5. Vettore
                  6. Elemento
                     7. Dato

Facciamo subito un esempio concreto per fissare le idee:

0. Comportamenti collettivi e associazionismo
   1. Conflitti di lavoro
      2. Ore scioperate
         3. Ore scioperate nell'industria per rivendicazioni salariali
            4. Ore annue espresse in migliaia
               5. Sezione spaziale per provincia, anno 1981
                  6. Provincia di Cuneo
                     7. Livello di affidabilità I (massimo)

Per percorrere con agilità gli anelli della catena che dal tema conduce fino al dato singolo è essenziale mettere a fuoco la natura degli operatori che permettono di passare da un livello all'altro. Alcuni di essi sono relativamente banali (passaggio da 0 a 1) o sono già stati analizzati dettagliatamente nel progetto di fattibilità (passaggio da 6 a 7). Altri sono più complessi e meritano una

descrizione più accurata (soprattutto i passaggi intermedi). Vediamoli dunque uno per uno.

1.1. Tema e sottotema

Si tratta di una normale operazione di selezione di un sottoinsieme da una lista di sottoinsiemi disgiunti (v. Appendice 1 del progetto di fattibilità). Qui è solo il caso di aggiungere due osservazioni:

a) la gerarchia delle aree e delle sottoaree tematiche riflette l'organizzazione fissa dell'archivio, che attribuisce univocamente un indicatore ad una sottoarea tematica, in modo da eliminare duplicazioni e ridondanza;
b) a regime l'utente non sarà vincolato a seguire la rappresentazione interna del sistema, ma potrà formulare il tema della sua richiesta direttamente in linguaggio naturale, affidando al sistema stesso la sintesi di un "dominio di ricerca", ossia di un mix di aree e/o sottoaree tematiche congruente con la sua richiesta.

In concreto questo vuol dire che un modulo del sistema sarà costituito da una rete semantica capace di collegare fra loro parole-chiave, e sintetizzare un mix equilibrato (nè troppo grande nè troppo piccolo) di aree e/o sottoaree tematiche candidate a soddisfare la richiesta dell'utente. Naturalmente la costruzione del "dominio di ricerca" ottimale non può essere affidata interamente al sistema, ma richiede l'apertura di un ciclo di interazioni in cui la capacità di proposta del sistema viene integrata dalla capacità di selezione e di controproposta dell'utente.

1.2. Il concetto di indicatore sorgente

Ogni sottoarea tematica può essere concepita come una lista di indicatori sorgente. Ma che cosa è un indicatore sorgente, e che cosa lo differenzia dagli indicatori veri e propri?

La differenza può essere illustrata con un esempio. Il numero totale di ore scioperate, indipendentemente dalla causa, dal settore produttivo, dal livello del conflitto etc. è un indicatore sorgente. Il numero di ore scioperate nel settore metalmeccanico per il rinnovo del contratto di lavoro è un indicatore vero e proprio. Più in generale si può proporre la seguente definizione:

Indicatore sorgente. Dato un dominio statistico omogeneo si chiama indicatore sorgente qualsiasi indicatore non ulteriormente aggregabile nell'ambito del dominio.

Così, restando al dominio statistico "conflitti di lavoro", gli indicatori sorgente possibili sono solo tre: numero totale di conflitti, numero totale di lavoratori partecipanti, numero totale di ore scioperate. Essi individuano indicatori sorgente differenti perchè non possono essere sommati fra loro, nè consentono ulteriori aggregazioni.

Ques'ultima condizione vale, naturalmente, non in senso logico e assoluto, ma relativamente ad un ambito di dati reciprocamente confrontabili. Se disponessimo, ad esempio, di una statistica delle ore di lavoro confrontabile con quella delle ore di sciopero nulla vieterebbe di aggregare fra loro i due tipi di allocazione del tempo e considerare quindi come indicatore sorgente la somma di entrambe. Questa operazione non è illegittima in sè, ma lo diventa dato il tipo di informazione statistica disponibile. Poichè gli orari di lavoro sono valutati da organi e con modelli di rilevazione difformi e inconfrontabili con quelli relativi agli scioperi, orari di lavoro e conflitti di lavoro costituiscono due domini statistici distinti, e dunque la identificazione degli indicatori sorgente deve procedere separatamente nei due casi.

1.3. Lo schema di disaggregazione

Il termine indicatore sorgente è stato scelto per sottolineare il fatto che esso rappresenta la radice, la fonte, o la "sorgente" appunto, dell'insieme di indicatori veri e propri che da esso si possono ricavare per disaggregazione. Detto in altri termini il nucleo del passaggio da un indicatore sorgente alla famiglia degli indicatori da esso generati è costituito dal suo schema di disaggregazione.

Uno schema di disaggregazione può essere visto come un insieme di suddivisioni distinte ma fra loro ricombinabili del medesimo insieme fondamentale.

Restando all'esempio delle ore di sciopero, un possibile shema di disaggregazione è il seguente:

Criterio I - Ramo produttivo

1. Agricoltura
2. Industrie estrattive

.....

10. Pubblica amministrazione

Criterio II - Causa del conflitto

1. Rivendicazione salariale
2. Licenziamento
3. Altra causa

Criterio III - Livello del conflitto

1. Di azienda
2. Di categoria
3. Di più categorie.

Ogni criterio -o "principium divisionis"- dà luogo ad un certo numero di nodi, ciascuno dei quali rappresenta una determinata "fetta" della totalità designata dall'indicatore sorgente. L'identificazione di un indicatore vero e proprio avviene specificando quali criteri si desiderano utilizzare e selezionando, per ogni criterio specificato, uno dei nodi disponibili. In questo senso si può dire che un indicatore vero e proprio è una lista di nodi, uno per ogni criterio dello schema di disaggregazione (ai criteri non specificati perchè considerati irrilevanti viene automaticamente attribuito il "nodo zero", che equivale a selezionare la totalità dei soggetti secondo il criterio omesso).

Anche nell'ambito del medesimo criterio i nodi dello schema di disaggregazione non si riferiscono necessariamente ad insiemi disgiunti (schema di disaggregazione come partizione) nè ad insiemi gerarchicamente ordinati (schema di disaggregazione come albero). Entrambe queste alternative, infatti, sono incompatibili con la frammentarietà e la discontinuità dell'informazione statistica disponibile. Per chiarire questo punto riprendiamo il consueto esempio delle ore di sciopero. Nel periodo 1949-1986 l'Istat ha utilizzato almeno cinque modi diversi di suddividere le cause dei conflitti di lavoro. Se non vogliamo perdere nessuna delle possibilità di aggregazione effettivamente presenti nei dati siamo costretti a rappresentare i nodi del criterio "Causa del conflitto" nel modo seguente:

0. Conflitti
    1. Estranei al rapporto di lavoro
    2. Originati dal rapporto di lavoro
        3. Rinnovo del contratto
        4. Solidarietà
        5. Licenziamenti e sospensioni
            6. Licenziamenti
            7. Sospensioni
        8. Rivendicazioni salariali, economiche e normative
            9. Rivendicazioni salariali
            10. Rivendicazioni economiche e normative

11. Altre cause
12. Mancato pagamento spettanze arretrate
13. Inadempienza contrattuale o pluralità di cause

Aggregazioni speciali:

14. Altra causa: 4+7+10+11
15. Altra causa: 1+3+4+7+10+12+13
16. Altra causa: 1+4+7+10+12+13

In casi come questo sia la scelta di adottare la partizione più "fine" (i nove insiemi disgiunti ed esaustivi più elementari: 1,3,4,6,7,9,10,12,13) sia la scelta di adottare uno schema gerarchico unico (insiemi 1-13) comportano la rinuncia alla maggior parte dell'informazione statistica disponibile. Quest'ultima non è infatti nè completa, nè dotata di una struttura costante nel tempo. Se non si vuole ometterne una parte la lista dei nodi deve contenere tutte le aggregazioni effettivamente impiegate dall'Istat, comprese quelle non integrabili in uno schema gerarchico unico ("Aggegazioni speciali").

1.4. Indicatori e varianti

La distinzione fra indicatore e variante è già stata ampiamente discussa nel progetto di fattibilità (paragrafo 2 della parte I). Qui è soltanto il caso di ricordare che mentre la cadenza della rilevazione (mensile, annuale etc.) è uno degli elementi che permettono di differenziare le varianti del medesimo indicatore (la serie mensile e la serie annuale delle ore scioperate nell'industria costituiscono due varianti distinte) il livello di aggregazione territoriale dei dati (comunale, provinciale, regionale, nazionale) non ha il medesimo ruolo. Questa differenza nelle funzioni dello spazio e del tempo è una conseguenza della natura stessa dell'Archivio, che è concepito innanzitutto come collettore di serie storiche, e solo mediatamente come fascio di "sezioni" spaziali

(cross-section).

Per rovesciare la prospettiva di rappresentazione dell'Archivio, e vedere in cross-section ciò che è concepito essenzialmente in time series, occorre considerare l'anello successivo della catena, quello che collega varianti e vettori.

1.5. Dalla variante al dato

Una volta selezionata una variante di un indicatore (ex: scioperi annuali nell'industria) l'utente è giunto all'estremo confine di un mondo, quello dei concetti, e sta per entrare nel dominio di un mondo differente, quello dei dati. Qui gli si aprono essenzialmente due vie, a seconda che il suo obiettivo sia quello di conoscere un singolo dato ("Quante sono state le ore di sciopero nell'industria a Cuneo nel 1981?") oppure quello di ottenere un vettore, ossia un'intera distribuzione o seguenza di dati.

Nel primo caso egli dovrà selezionare l'unità di tempo (1981) e di spazio (provincia di Cuneo) cui il dato si riferisce, nonchè un determinato livello di affidabilità (v. par. 4.2., parte I, del progetto di fattibilità).

Nel secondo caso dovrà selezionare o l'unità di tempo (se desidera un vettore in cross-section) o l'unità di spazio (se desidera un vettore in time series), oltrechè, naturalmente, il livello di affidabilità dei dati.

E' il caso di sottolineare che l'organizzazione tridimensionale (spazio x tempo x livello di affidabilità) delle "matrici di accoglimento" dei dati fa sì che la semplice scelta di un elemento di un determinato vettore di dati non dia ancora accesso al dato singolo. Solo la specificazione dei livello di affidabilità desiderato permette di compiere l'ultimo passaggio, dal singolo elemento del vettore (che può essere anche una terna di dati) al dato vero e proprio.

2. LA GERARCHIA DELLE TRASFORMAZIONI ELEMENTARI

Il quadro delle statistiche ufficiali del dopoguerra si presenta come un immenso colabrodo. Indicatori disponibili per un certo periodo cessano di esserlo in quello successivo per riapparire, magari in forma leggermente modificata, in un periodo ancora successivo. Indicatori disponibili su base nazionale non lo sono su base regionale, o lo sono solo per certi sottoperiodi. Altri indicatori, magari disponibili fin dai primi anni del dopoguerra, cessano improvvisamente e misteriosamente di esistere in un dato anno per riemergere, come fiumi carsici sotteranei, in un periodo talora anche lontano senza alcuna plausibile spiegazione.

In queste condizioni una delle funzioni-chiave dell'Archivio diventa quella di mettere a disposizione dell'utente non solo i dati esistenti e una mappa dettagliata delle loro lacune, ma anche gli strumenti per colmarle.

Il problema delle trasformazioni cui i dati possono essere sottoposti è già stato in parte affrontato nel progetto di fattibilità mediante il concetto di trasformazione elementare (par. 2 della parte I). Per trasformazione elementare si intende, in generale, una trasformazione dei dati effettuata mediante modelli statistici relativamente semplici, o poveri di teoria. I casi più banali sono costituiti dalle dilatazioni a coefficienti noti (raccordo di due varianti espresse in unità di misura diverse), e dalle operazioni di aggregazione temporale e spaziale dei dati (passaggio da dati mensili a dati annuali, da dati provinciali a dati regionali etc.). A un livello intermedio di complessità abbiamo il raccordo di due varianti di un medesimo indicatore in condizioni di non consocenza dei parametri di raccordo (ex: serie trimestrali delle forze di lavoro fra l'ottobre del 1976 e il gennaio del 1977). Al livello più alto di complessità troviamo le operazioni di disaggregazione temporale e spaziale dei dati ("stagionalizzazione" di dati annui, stima di dati provinciali a partire da dati regionali etc.). In casi come questi la possibilità di scendere la scala di aggregazione dei dati si basa su due circostanze:

a) la disponibilità di dati del livello di aggregazione inferiore per periodi diversi da quello considerto;
b) la presenza e l'identificabilità di strutture di autocorrelazione temporale o spaziale dei dati.

L'accostamento fra il caso della disaggregazione temporale e quello della disaggregazione spaziale può sembrare arbitrario ma si basa in realtà sulla convergenza fra i risultati di due discipline ausiliarie della statistica: l'econometria, tradizionalmente orientata alle strutture di autocorrelazione temporale, e la demografia, sempre più attenta al problema delle strutture di autocorrelazione spaziale.

Poichè il livello di affidabilità dei risultati dei vari tipi di trasformazione elementare non è costante ma diminuisce man mano che cresce il numero di assunzioni implicite nella trasformazione conviene ordinare le trasformazioni stesse in una gerarchia, dalle più sicure alle più incerte:

Gerarchia delle trasformazioni elementari

Trasformazioni di classe 1:
- aggregazione temporale
- aggregazione spaziale
- dilatazioni a coefficienti noti

Trasformazioni di classe 2:
- raccordo di serie storiche mediante trasformazioni affini
- stima di dati definitivi mediante dati provvisori

Trasformazioni di classe 3:
- disaggregazione temporale
- disaggregazione spaziale

Questi tipi di trasformazioni dovrebbero essere sempre a disposizione dell'utente, e consentirgli la ricucitura automatica della maggior parte delle configurazioni di dati lacunose.

Al di là di questi tipi di trasformazioni, che abbiamo definito

elementari proprio perchè non richiedono l'intervento di assunzioni teoriche specifiche, inizia il dominio delle trasformazioni complesse, "ricche di teoria", come le interpolazioni, la stima di indicatori lacunosi mediante altri indicatori, la trasformazione di insiemi di indicatori semplici in indici più o meno sofisticati. Questo secondo tipo di trasformazioni richiedono assunzioni forti da parte dell'utente, e non possono quindi essere codificate in modo generale, mediante schemi di manipolazione dei dati definiti a priori. In questi casi il sistema si limiterà a fornire una serie di utilities, lasciando all'utente la responsabilità di combinarle e integrarle in uno schema unitario e teoricamente plausibile.

3. UNA TIPOLOGIA DELLE RICHIESTE DELL'UTENTE

L'utente che si accosta all'Archivio degli indicatori sociali può averne una conoscenza più o meno dettagliata e completa, così come può avere degli obiettivi più o meno specifici. Già queste due dimensioni dànno luogo a quattro "tipi ideali" di interazione sistema-utente, che possiamo cercare di concretizzare con quattro esempi di richieste:

		CONOSCENZA ARCHIVIO: VAGA	CONOSCENZA ARCHIVIO: PRECISA
OBIETTIVI: GENERALI		TIPO A "Che cosa c'è nell'Archivio?"	TIPO B "Vorrei un data set SAS 1946-1986 con tutte le serie storiche annue"
OBIETTIVI: SPECIFICI		TIPO C "Esiste il dato delle ore di sciopero nell'aprile del 1963"?	TIPO D "Vorrei la serie storica 1946-1986 delle ore di sciopero annue in Piemonte"

Questa classificazione delle richieste si complica ulteriormente se teniamo conto della qualità dei dati che l'utente è disposto a prendere in considerazione.

Questo non solo perchè nell'Archivio il dato non è una grandezza scalare ma è un vettore a tre posti non necessariamente completo (dato provvisorio, dato semidefinitivo, dato definitivo), ma perchè in molti casi i dati richiesti possono non esistere come dati grezzi ma essere producibili mediante trasformazioni elementari di dati grezzi (dati virtuali).

In casi come questi il sistema dovrebbe, in un certo senso, rispondere alle richieste dell'utente introducendo più o meno sottili "distinguo":
"Sì, il dato che cerchi c'è ma non supera il livello II di affidabilità"
oppure:
"No, il dato che cerchi non esiste, però posso generarlo con una trasformazione elementare di classe 3".

La varietà delle possibili richieste dell'utente cresce poi ancora di più se, oltre ai data base delle serie originarie e delle (eventuali) serie derivate generate mediante trasformazioni elementari, si prende in considerazione il data base delle serie derivate generate mediante trasformazioni non elementari, (su questo punto vedi il progetto di fattibilità, paragrafo 5.2. della parte I, e Appendice 5).

Di fronte a questa moltiplicazione delle possibilità di interazione sistema-utente è essenziale, prima ancora di definire specifici percorsi di discesa lungo l'albero delle possibilità, fissare alcuni principi generali che dovrebbero orientare la costruzione di tali percorsi.

Il primo principio è quello della delimitazione spazio-temporale. Quale che sia il livello di informazione dell'utente e quali che siano i suoi obiettivi, fra le prime richieste che il sistema gli farà vi è quella di precisare l'ambito storico e geografico della sua ricerca di informazioni.

Il secondo principio è quello della anticipazione dell'output. Per guidare efficacemente l'utente nella scelta dei menù il sistema ha bisogno di sapere a che "tipo ideale" la richiesta dell'utente appartiene. Più precisamente le risposte finali alle richieste di un utente possono assumere quattro forme-base.

1. Una o più liste di nomi di indicatori, e relative varianti.
2. Un data set SAS, o in time series o in crossection, contenente un certo numero di vettori.
3. Un vettore di dati (serie storica o sezione spaziale).

4. Un dato singolo, eventualmente replicato ai vari livelli di affidabilita'

Il terzo ed ultimo principio riguarda la qualità dei dati. L'Archivio degli indicatori sociali contiene sette diversi tipi di dati:

Dati grezzi: definitivi
semidefinitivi
provvisori

Dati virtuali: di classe 1
di classe 2
di classe 3

Stime e indici.

Ciò rende enormemente più complicato il senso di domande apparentemente innocenti come questa:
"Esiste la serie trimestrale 1959-1986 degli occupati nell'industria in provincia di Torino?".

La risposta del sistema può variare notevolmente a seconda che l'utente:

a) si riferisca alla serie completa o sia anche disposto a considerare una serie lacunosa;
b) accetti o meno di sottoporre i dati a trasformazioni elementari di classe 2 per raccordare tratti non omogenei (cambiamenti nella definizione di industria o nella definizione di occupato);
c) accetti o meno di sottoporre i dati a trasformazioni elementari di classe 3 per stimare il dato torinese negli anni in cui l'Istat riporta solo il dato piemontese.

Di qui la necessità di fissare in anticipo, ossia prima di inoltrarsi nel labirinto delle alternative e delle domande, quali sono gli standard qualitativi delle richieste dell'utente, sia che queste siano mere richieste di informazione ("quali serie storiche esistono

sul tema "X"?) sia che queste siano vere e proprie richieste di accesso ai dati.

Quest'ultimo problema, quello del rapporto fra dati grezzi, o reali, e dati generati, o virtuali, permette di illustrare in modo particolarmente efficace il carattere critico e "costruttivista" della concezione dei dati implicita nella architettura generale dell'Archivio. Critico perchè il dato non viene mai considerato come una realtà ultima ed univoca ma, letteralmente, come un vettore di possibilità su cui ragionare e fra cui scegliere. Costruttivista perchè, distinguendo fra varianti diverse del medesimo indicatore, fra dati reali e dati virtuali, fra trasformazioni elementari e trasformazioni non standard, tende a sottolineare quanto ricco e complesso sia il lavoro di selezione, trasformazione ed elaborazione di cui i cosiddetti "fatti sociali" sono il risultato finale.

BIBLIOGRAFIA

Bauer R.A. (ed.),
1966 Social Indicators, Cambridge, Mit Press.

Belnap N.D. - Steel T.B.,
1976 The Logic of Questions and Answers, Yale University Press.

Cartocci R.,
1984 Concetti e indicatori: il contributo della nuova retorica, in "Sociologia e ricerca sociale", V, 13.

Coombs C.H.,
1964 A Theory of Data, New York, Wiley.

CSI-Piemonte,
1985 Un'interfaccia per l'interrogazione in linguaggio naturale.

Curatolo R.,
1979 Indicatori sociali per la Toscana, Irpet, Firenze.

Dodd S.C.,
1940 Dimensions of Society, New York, xxxx.

Galtung J.,
1967 Theory and methods of Social Research, Oslo, Universitetvorlaget

Gerbner F.,
1969 Toward "Cultural Indicators": the Analysis of Mass Mediated Message System, in "Audiovisual Communication Review", 1969, 2.

Lazarsfeld P.A.,
1958 Evidence and Inference in Social Research, in D. Lerner (ed.)

Evidence and Inference, New York, Free Press.

Marradi A.,

1984 Concetti e metodi per la ricerca sociale, Firenze, La Giuntina

1987 La validitè des indicateurs et la fidelitè de donnees, in corso di pubblicazione.

Marradi A. (a cura di),

1988 Costruire il dato, Milano, Franco Angeli.

Niceforo A.,

1921 Les indices numèriques de la civilization et du progrès, Paris, Flammarion.

Odland J.,

1987 Spatial Autocorrelation, Sage.

PARTE III IL PROTOTIPO

1. SCOPI DEL PROTOTIPO

La traduzione in un prodotto concreto e funzionante (anche se a livello di prototipo) del disegno teorico dell'archivio tracciato nelle pagine precedenti costituisce senza dubbio un progetto ambizioso sia per la complessità e varietà dell'argomento (la "catena di concetti" che dal tema conducono al dato) sia per le difficoltà tecniche e organizzative (scelte e vincoli posti dall'utilizzo di specifici prodotti hardware e software, reperimento e documentazione di una sufficiente quantità di dati).

L'obiettivo primario che ha determinato la decisione di realizzare il "Prototipo di sistema per la documentazione e reperimento degli indicatori socio-demografici" è stato certamente quello di dimostrarne la fattibilità. Abbiamo, in altre parole, voluto accertarci della congruenza interna e della tenuta esplicativa dell'insieme dei concetti elaborati nella fase di progettazione logico-funzionale.

Questo obiettivo ci sembra raggiunto con piena soddisfazione in quanto al di là delle nostre stesse aspettative il prodotto software che nel seguito verrà illustrato non dispone soltanto di un ambiente di ricerca in grado di far "conoscere" cosa c'è nell'archivio ma anche di un primo embrione di ambiente di analisi dei dati capace di offrire una o più matrici dati contenenti gli indicatori selezionati dall'utente nella fase di ricerca.

La fase di realizzazione del prototipo è stata caratterizzata da una intensa interazione con il disegno teorico di partenza costringendoci spesso a rivedere e affinare l'apparato metodologico e concettuale. E' interessante notare come questo "fedback" non sia stato generato, se non in maniera marginale, da aspetti tecnici legati ai vincoli posti dal software o dall'hardware utilizzati quanto piuttosto dalla necessità di rendere l'interazione uomo-macchina il più chiara e precisa possibile. Questo processo che ha interessato con successive ridefinizioni e precisazioni concetti come "INDICATORE" "INDICATORE SORGENTE" "VARIANTE" "SCHEMA DI DISAGGREGAZIONE", da un lato ha permesso l'eliminazione di soluzioni di continuità tra l'architettura logica dell'archivio e le operazioni di ricerca ordinate dall'utente,

dall'altro ha consentito di sottoporre l'apparato concettuale ad una costante verifica di "tenuta" in condizioni concrete di utilizzo. Questa reciproca influenza tra architettura logica, apparato concettuale e sua realizzazione pratica è da tenere presente anche in previsione della messa a punto di uno strumento non più solo prototipale.

Come suggerisce la saggezza popolare infatti "Fra il dire e il fare c'è di mezzo il mare" ma non bisogna dimenticare che soprattutto nella realizzazione di progetti come questo il mare delle difficoltà da superare va attraversato più volte accettando il fatto che l'opera di realizzazione pratica sveli ambiguità e imprecisioni suggerendo la revisione dell'apparato teorico, dell'idea di partenza, che è in questo modo soggetto ad una costante opera di adattamento.

Se le ambiguità e le imprecisioni dell'apparato concettuale possono essere messe a dura prova e quindi verificate durante la realizzazione pratica del disegno teorico, non altrettanto si può dire delle ambiguità, imprecisioni, elementi di incomprensione che possono venire introdotti nelle forme che assume l'interazione tra il sistema uomo e il sistema macchina. Durante la fase di programmazione dei cicli domanda-risposta che determinano appunto le forme di tale interazione è necessario ottimizzare e sintetizzare l'informazione che passa da un sistema all'altro. L'informazione diretta al sistema macchina è certamente sintetica e non ambigua (avendo escluso l'utilizzo del linguaggio naturale tale informazione si concretizza quasi sempre in una cifra, una lettera o un tasto funzionale) non altrettanto accade nel caso dell'informazione che la macchina fornisce all'utente affinchè possa operare le sue scelte. In questo caso il linguaggio naturale è utilizzato (questa è stata la nostra scelta) ma l'ambiguità e l'incomprensione insita nelle frasi o nelle videate proposte dalla macchina all'utente possono riemergere come elementi che disturbano la semplicità di utilizzo del sistema.

Un ulteriore obiettivo raggiunto con la realizzazione del prototipo è così quello di consentire una verifica ed una eventuale messa a punto delle forme di interazione uomo-macchina. Verifica e messa a punto che

può avvenire soltanto dopo un utilizzo del prototipo allargato a più utenti, anche differenti per esigenze, atteggiamento culturale, consocenze statistiche e metodologiche nell'uso dei dati socio-demografici.

L'IRES può costituire un ottimo ambiente per tale verifica grazie alle caratteristiche peculiari dell'attività svolta, alla quantità e qualità del patrimonio di informazioni ed esperienza maturati nella sua storia.

Infine, l'ultimo scopo (non certo per importanza), che la realizzazione del prototipo consente è quello di disporre, fin da subito di un embrione di sistema strutturato di documentazione del patrimonio dati dell'Istituto in grado di evitare o abbreviare la fase di reperimento dati che richiede sempre un notevole dispendio di energie e di tempo. Dispendio di risorse ed energie che è tanto più grande quanto più, come nel caso degli indicatori socio-demografici, l'insieme dei dati si presenta ancora attualmente alquanto frammentario, disperso ed eterogeneo.

2. LIMITI E SPECIFICITA'

Un'idea forte ha guidato le scelte fatte nella realizzazione del prototipo: "l'uso dello strumento informatico deve essere semplice".

Questo imperativo si concretizza in tre aspetti:

a) l'utente deve poter istruire la macchina senza dover tenere a mente termini o comandi di alcun tipo, ed inoltre le conseguenze di ogni specifica scelta dell'utente devono essere spiegate esaustivamente e in modo chiaro.

b) L'utente deve poter sapere in ogni momento in quale stadio di "ricerca" o "analisi dati" si trova e cosa può o non può fare a partire da quello stadio.

c) Le informazioni fornite dalla macchina in risposta ai comandi dell'utente devono rispecchiare e ulitizzare soltanto l'impostazione logica e i caratteri dell'archivio senza appesantire il colloquio con riferimenti alla "struttura fisica" dei dati (nomi di files, vettori ecc.).

L'utilizzo del "linguaggio naturale" può sembrare, a prima vista, un ottimo strumento in grado di garantire in maniera ottimale il raggiungimento dei nostri obiettivi. Tuttavia alcune limitazioni tipiche dei sistemi capaci di dialogare con l'utente accettando richieste in linguaggio naturale vanificherebbero il progetto stesso di archivio (1). Fra queste possiamo ricordare:

- competenza linguistica su domini ristretti;
- difficoltà di gestire il cambiamento di ambiente, ovvero di cogliere il "contesto" delle richieste;
- allungamento dei tempi di attesa dovuto alla pesantezza e alla complessità delle operazioni di parsing, interpretazione, ecc.;
- elevati tempi di progettazione e sperimentazione.

Inoltre facendo specifico riferimento ad un colloquio uomo-macchina volto ad accertare l'esistenza e la disponibilità di dati statistici

---

(1) Si veda a questo proposito: L.I.A. "Un'interfaccia per l'interrogazione in linguaggio naturale di un data-base" CSI-PIEMONTE, 1985.

sembra possibile ridurre il ventaglio delle domande proponibili alla macchina in un numero relativamente limitato di "TIPI di domanda" sufficientemente esaustivo (1).

Sulla base di queste considerazioni la scelta di un tipo di interazione guidata da menù (pannelli) è sembrata una scelta quasi obbligata.

Nell'ambito di questa scelta è stato fatto un notevole sforzo per adottare un formato il più possibile uniforme tra i diversi pannelli in modo da ridurre al minimo le difficoltà di adattamento dell'utente nelle diverse fasi dell'interazione.

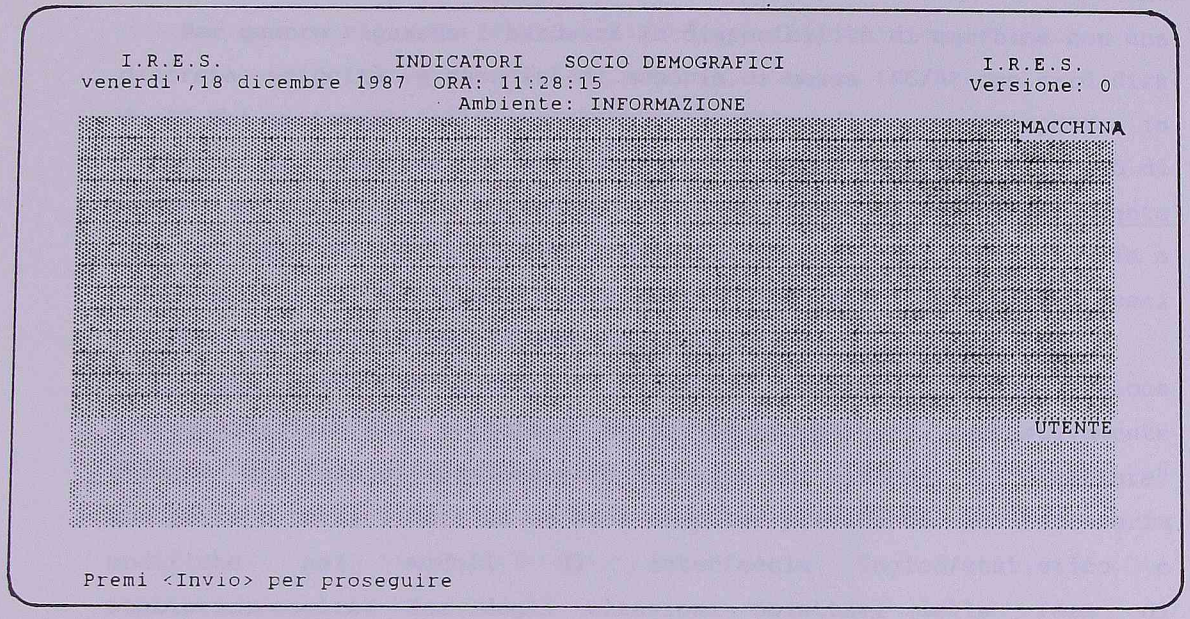
Con l'ausilio di un video a colori tutti i menù dell'ambiente di informazione sono stati suddivisi in tre aree (v. Fig. 1).

1) Area bianca (in alto e in basso rispetto al video) viene utilizzata per informazioni di contesto (ambiente di riferimento, tasti da premere, informazioni sulle operazioni di ricerca compiute, ecc.).
2) Area verde (tratteggio scuro sulla figura) contrassegnata dalla scritta "MACCHINA", in alto a destra. E' la zona del video deputata a rappresentare il flusso di informazioni che dalla macchina sono destinate all'utente (operazioni che la macchina è in grado in quel momento di eseguire, risultati di operazioni svolte, ecc.). Tale area viene espansa quasi ad occupare lo schermo interno quando l'utente non può fare altro che "attendere" o quando la risposta della macchina prevede un ritorno obbligato ad una precisa fase del colloquio (per es. quando si richiedono informazioni dettagliate (TASTO "H") su una delle possibili scelte proposte dalla macchina prima che l'utente prenda la sua decisione).
3) Area blù (Tratteggio chiaro sulla figura) contrassegnata dalla scritta "UTENTE" in alto a destra. Questa zona del video è riservata a rappresentare le scelte, i comandi che l'utente impone alla macchina.

---

(1) N.D. BELNAP, T.B. STEEL: "The logig of questions and answers" Yale University Press, 1976.

Fig. 1


I.R.E.S. INDICATORI SOCIO DEMOGRAFICI I.R.E.S.
venerdi',18 dicembre 1987 ORA: 11:28:15 Versione: 0
Ambiente: INFORMAZIONE
MACCHINA
UTENTE
Premi <Invio> per proseguire

3. HARDWARE E SOFTWARE

Considerazioni legate alla semplicità d'uso, flessibilità e disponibilità hanno infine costituito gli elementi decisivi nella scelta dell'ambiente hardware e del software da utilizzare per la realizzazione del prototipo.

Per quanto riguarda l'hardware la disponibilità di macchine con una discreta velocità e capacità di memoria di massa (PC/AT con hard disk da 30 Mb) unitamente alla imminente prospettiva di poter disporre in Istituto di macchine di gran lunga più potenti (32 biT con 200 Mb di memoria su disco) hanno consentito di adottare una soluzione totalmente locale. Sia le basi di dati statistici (i dati veri e propri) sia i files contenenti le "descrizioni" degli indicatori e dei dati stessi possono ragionevolmente risiedere su un'unità locale.

Questa soluzione ha avuto il pregio di rendere più snella e veloce la realizzazione del prototipo senza compromettere, successivamente durante una fase di funzionamento a regime del sistema, di "dirottare" una parte o tutti i dati su un host computer prevedendo le necessarie modifiche nei moduli di interfaccia logico/statistico e statistico/logico. Uno degli ulteriori obiettivi della fase di sperimentazione del funzionamento dell'archivio tramite il prototipo potrebbe essere proprio quello di definire l'insieme degli indicatori e dei dati che vengono "consultati" più sovente, utilizzando così questo criterio empirico per stabilire il luogo fisico di residenza degli stessi.

Relativamente al software operando nell'ambito del sistema operativo MS-DOS le scelte sono cadute sul package SAS (Statistical Analysis System) che, ora anche su Personal Computer, con il proprio sistema di interfaccia utente (DMS-data Manages system) costituisce un ambiente di analisi dati che si incunea perfettamente nell'architettura globale del sistema di documentazione e reperimento degli indicatori. Altre caratteristiche del SAS, già citate in altra parte del presente volume, lo rendono particolarmente adatto ai nostri scopi: possiamo ricordare le due che più lo differenziano da altri prodotti analoghi:

- elevate capacità di programmazione

- incorporazione del linguaggio IML (operazioni algebriche su matrici).

Se la scelta dell'ambiente statistico appare in una certa misura "definitiva" non altrettanto si può dire per lo strumento software utilizzato come supporto per la programmazione dell'interfaccia utente e delle operazioni di ricerca e recupero delle informazioni sugli indicatori.

La scelta è caduta su dBASE III e sul compilatore CLIPPER due strumenti che per manegevolezza, e flessibilità sono certamente all'avanguardia.

Grazie alle possibilità di programmazione offerte da questo software è stato possibile realizzare in tempi abbastanza brevi un'interfaccia utente rispondente ai requisiti preposti per il prototipo, così come è stato possibile realizzare un primo nucleo di funzioni che consentono di percorrere un "sentiero di ricerca" già discretamente complesso.

A regime tuttavia il sistema presenterà caratteristiche che rischiano di rendere inutilizzabile questo tipo di software; in particolare:

- la complessità degli "alberi di ricerca" e dei percorsi che il sistema dovrà essere abilitato a compiere durante la fase di ricerca tenderà a crescere notevolmente.
- Le "conoscenze" sul dominio dei dati memorizzati in possesso della macchina sono ora solo parzialmente separate dai singoli algoritmi di calcolo logico. In altre parole molte informazioni che istruiscono la macchina su come deve comportarsi in presenza di diverse situazioni sono al momento integrate nel flusso dei programmi. Questo fatto che come si sà è connaturato all'utilizzo di linguaggi imperativi costituisce un grave handicap in ogni successiva fase di aggiornamento e/o incremento della base dati.

Questi due inconvenienti possono essere superati facendo ricorso a linguaggi di programmazione logica (LISP, PROLOG, ecc.) che per loro natura sono particolarmente adatti a descrivere e percorrere alberi di ricerca anche assai complessi. Inoltre la separazione consentita e in un certo senso imposta dall'uso di questi linguaggi tra "base di conoscenza" e "regole" logiche garantisce la possibilità di effettuare

agevolmente cambiamenti e modifiche.

Avendo tuttavia sperimentato nella fase prototipale la varietà e la diversificazione delle esigenze che la realizzazione di un tale sistema comporta, ci sembra importante sottolineare la necessità di effettuare a questo proposito scelte non esclusive.

Così come dBase e CLIPPER si sono dimostrati ottimi strumenti in particolare per la realizzazione dell'interfaccia utente, così riteniamo particolarmente promettente l'uso di linguaggi logici e di strumenti di intelligenza artificiale come LISP nell'ampliamento e potenziamento delle funzioni dell'ambiente di ricerca, soprattutto per consentire all'utente di effettuare richieste più "sfumate", e imprecise dotando la macchina delle competenze necessarie per instaurare con l'utente un dialogo appropriato.

Questo obiettivo ci sembra raggiungibile senza tuttavia cadere negli inconvenienti già ricordati relativi all'uso del linguaggio naturale. Il nostro approccio intende in sostanza superare la semplicistica equazione: intelligenza artificiale = linguaggio naturale cercando invece di porre le potenzialità offerte dai linguaggi usati in intelligenza artificiale al servizio di un'interfaccia per molti versi "tradizionale".

Il prototipo di sistema già realizzato può essere così visto come un primo nucleo rispetto al quale è possibile procedere a successivi potenziamenti e ampliamenti tramite l'integrazione di moduli (funzioni e procedure) realizzati sfruttando gli strumenti software più adeguati allo specifico scopo.

4. FUNZIONAMENTO

Entriamo ora maggiormente nel merito delle caratteristiche del prototipo del sistema per la documentazione e reperimento degli indicatori socio demografici evidenziando oltre alle funzioni che è al momento in grado di svolgere anche le riduzioni e semplificazioni che è stato necessario adottare.

All'accensione della macchina il sistema si presenta all'utente ponendosi direttamente nell'ambiente di: INFORMAZIONE.

L'accesso all'ambiente di ANALISI DATI verrà suggerito e consentito in seguito qualora le operazioni di ricerca sugli indicatori disponibili abbiano avuto buon esito mentre l'accesso all'ambiente di ALIMENTAZIONE è "riservato" e non disponibile per il generico utente.

Nell'ambiente di alimentazione deve essere svolta una serie di operazioni che, per la loro delicatezza richiedono, al momento, l'utilizzo di personale specificatamente addestrato.

Nell'ambiente di informazione la macchina si presenta all'utente comunicandogli quali sono i tipi di DOMANDE a cui è in grado di rispondere (fig. p1 in APPENDICE). I tre tipi di domande presentate sono indicative delle capacità di cui il sistema potrebbe essere dotato in una successiva versione.

Attualmente il prototipo è in grado di rispondere alla prima domanda che costituisce il metodo più analitico disponibile di ricerca e recupero di informazioni. Gli altri tipi di domande possono essere visti come "scorciatoie" di questo primo metodo. Se da un lato la caratteristica di queste ultime domande sarà quella di abbreviare i tempi di ricerca puntando direttamente l'attenzione sui dati di interesse dell'utente, dall'altro la peculiarità della prima domanda consiste nel consentire all'utente di farsi un'idea della struttura logico-formale dell'archivio.

Procedendo (selezione della domanda 1) la macchina propone all'utente di indicare le caratteristiche spaziali e temporali dei dati di suo interesse (Fig. p2). Se l'utente lo desidera può, rispondendo "NO" alla domanda, mettersi nella condizione di poter accedere a tutti i dati disponibili. Volendo invece ridurne il campo di ricerca può

specificarne i limiti SPAZIO-TEMPORALI (Fig. p3 e p4).

La macchina può ora procedere ad un primo processo di ricerca e selezione delle informazioni (Fig. p5) dopo aver chiesto la conferma sulle indicazioni fornite dall'utente (Fig. p4.1).

L'operazione di selezione dei dati pertinenti avviene a livello di Indicatori Sorgente. Individuato il sottoinsieme opportuno il sistema richiede la specificazione di un'AREA e una SOTTOAREA TEMATICA (Figg. p6, p6.1, p6.2, p6.3, da selezionare tra quelle che contengono indicatori sorgente pertinenti al sottoinsieme individuato (evidenziando queste ultime con una freccia "---->").

Nell'esempio cui fanno riferimento le figure in Appendice, non avendo posto alcuna limitazione spazio-temporale possiamo vedere il parco-dati di cui dispone attualmente il sistema.

L'utente può quindi indicare facilmente la sua scelta e la macchina restringe il campo degli indicatori sorgente disponibili a quelli pertinenti l'area tematica indicata.

Ora siamo in grado di vedere (Fig. p7) sullo schermo la lista degli indicatori sorgente nell'ambito di ciascuno dei quali l'utente può effettuare una ricognizione più approfondita fino ad arrivare alla selezione dei VETTORI desiderati.

Mentre informazioni dettagliate su ciascun indicatore sorgente possono essere ottenute premendo il tasto "H" (Fig. p7.1) con il tasto "X", premuto in corrispondenza di un qualsiasi elemento della lista è possibile procedere all'individuazione degli INDICATORI derivati dallo specifico indicatore sorgente. Tale operazione che può essere definita come una "discesa" lungo lo SCHEMA DI DISAGGREGAZIONE dell'indicatore sorgente si effettua selezionando uno alla volta i vari criteri di disaggregazione proposti dalla macchina (ovviamente differenti a seconda dell'indicatore sorgente in oggetto) e nell'ambito di ciascuno di questi selezionando il sottoinsieme desiderato (Fig. p8 e p8.1).

Nell'esempio riportato nel presente lavoro riusciamo a verificare l'esistenza e a "recuperare" per un loro successivo utilizzo (in ambiente di "ANALISI DATI") i dati relativi al "Numero di studenti nelle scuole medie inferiori". Questa operazione può essere eseguita ricorsivamente fin che lo si desidera in modo da recuperare tutti gli

INDICATORI disponibili nell'ambito dell'INDICATORE SORGENTE selezionato. Abbandonando l'indicatore sorgente selezionato (Fig. p9) è possibile (fig. p10) ricominciare la "discesa" dello schema di disaggregazione selezionando un altro indicatore sorgente, oppure, dato che disponiamo già di un certo numero di vettori, possiamo decidere di passare all'ambiente di ANALISI DATI. In questo secondo caso la macchina si incarica di effettuare una serie di controlli tra cui di particolare rilevanza:

a) la disponibilità dei vettori selezionati in matrici CROSS-SECTION o TIME SERIES (rispettivamente con unità di analisi il TERRITORIO o il TEMPO);
b) l'omogeneità tra i diversi vettori selezionati del livello di disaggregazione territoriale (nel caso CROSS-SECTION) o della CADENZA (nel caso TIME-SERIES).

L'esito di questi controlli influenza ovviamente il tipo di matrici che in ambiente di analisi dati verranno messe a disposizione dell'utente (1). Successivamente la macchina informa l'utente sui risultati dei controlli (Fig. p11) e consente, premendo il tasto "F2" di ottenere a video o su stampante le informazioni dettagliate sui dati disponibili (Fig. p11.1, p11.2, p11.3). Procedendo ulteriormente il sistema si preoccupa di trasmettere le informazioni necessarie alla creazione della/e matrice/i dati al sistema di trattamento statistico (SAS) (Fig. p12, p12.1). Terminata la fase di "generazione della base dati" nella parte alta del video si vede scorrere una prima stampa dei dati disponibili (Fig. 12.2). L'utente è ora in grado, utilizzando le potenzialità del package SAS di manipolare, elaborare, utilizzare nell'ambito di procedure statistiche i dati a sua disposizione.

Terminata la fase di analisi dati il sistema rientra in ambiente "INFORMAZIONE" rendendosi nuovamente disponibile ad effettuare altre ricognizioni sul contesto dell'archivio (Fig. p13).

Nonostante la sua veste prototipale e gli innumerevoli aspetti che

---

(1) In una versione a regime del prototipo verrà concesso all'utente in presenza di disomogeneità tra i dati, di scegliere il tipo di matrice, così come il livello di disaggregazione territoriale e di cadenza. Attualmente in presenza di disomogeneità, il sistema genera tante matrici quanti sono gli elementi disomogenei presenti.

ancora necessitano una più precisa messa a punto e ottimizzazione, il sistema si caratterizza, fin da ora, per una buona efficienza.

Come si può vedere dalla serie dei pannelli (da Fig. p1 a Fig. p13 riportate in appendice) l'intera sequenza di ricerca, recupero, e organizzazione in ambiente di analisi statistiche di dati descritta in precedenza ha richiesto meno di dieci minuti. Un lasso di tempo talmente breve che, anche nelle migliori condizioni ipotizzabili, non sarebbe sufficiente a consentire all'utente di scrivere ed eseguire il programma di lettura e generazione delle matrici dati.

APPENDICE : UN ESEMPIO DI INTERAZIONE UOMO-MACCHINA

Fig. P1

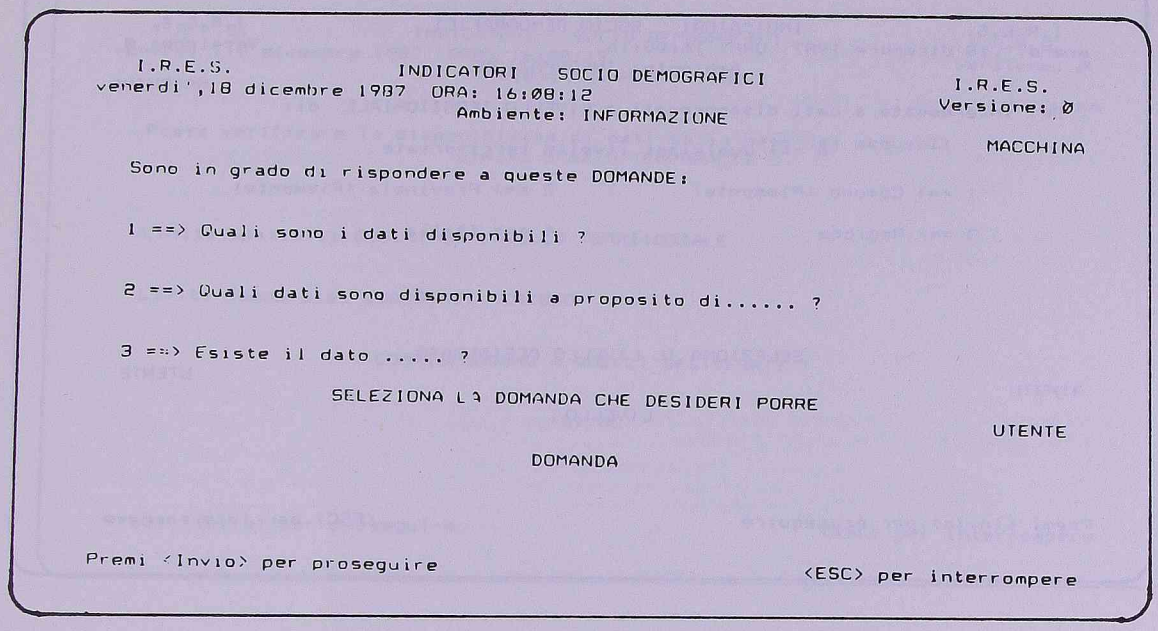

I.R.E.S.
INDICATORI SOCIO DEMOGRAFICI
venerdi',18 dicembre 1987 ORA: 16:08:12
Ambiente: INFORMAZIONE
I.R.E.S.
Versione: Ø
MACCHINA
Sono in grado di rispondere a queste DOMANDE:
1 ==> Quali sono i dati disponibili ?
2 ==> Quali dati sono disponibili a proposito di..... ?
3 ==> Esiste il dato ..... ?
SELEZIONA LA DOMANDA CHE DESIDERI PORRE
UTENTE
DOMANDA
Premi <Invio> per proseguire
<ESC> per interrompere

Fig. P2

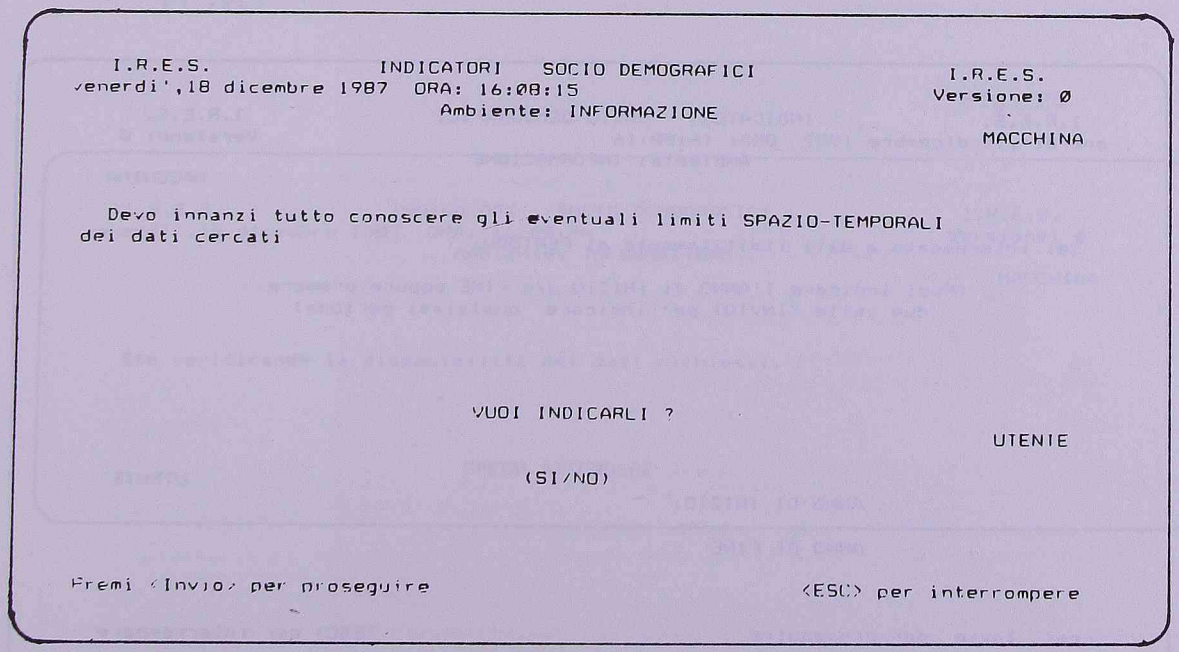

I.R.E.S.
INDICATORI SOCIO DEMOGRAFICI
venerdi',18 dicembre 1987 ORA: 16:08:15
Ambiente: INFORMAZIONE
I.R.E.S.
Versione: Ø
MACCHINA
Devo innanzi tutto conoscere gli eventuali limiti SPAZIO-TEMPORALI
dei dati cercati
VUOI INDICARLI ?
UTENTE
(SI/NO)
Premi <Invio> per proseguire
<ESC> per interrompere

Fig. P3

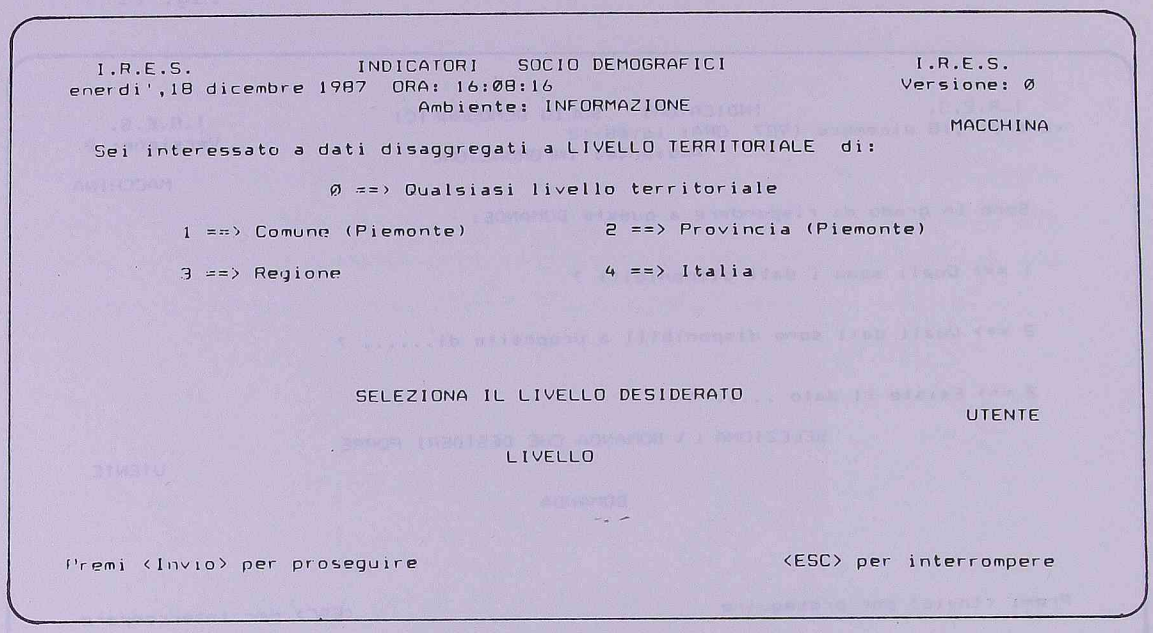

I.R.E.S. INDICATORI SOCIO DEMOGRAFICI I.R.E.S.
enerdi',18 dicembre 1987 ORA: 16:08:16 Versione: Ø
Ambiente: INFORMAZIONE
MACCHINA
Sei interessato a dati disaggregati a LIVELLO TERRITORIALE di:
Ø ==> Qualsiasi livello territoriale
1 ==> Comune (Piemonte)
2 ==> Provincia (Piemonte)
3 ==> Regione
4 ==> Italia
SELEZIONA IL LIVELLO DESIDERATO
UTENTE
LIVELLO
Premi <Invio> per proseguire
<ESC> per interrompere

Fig. P4

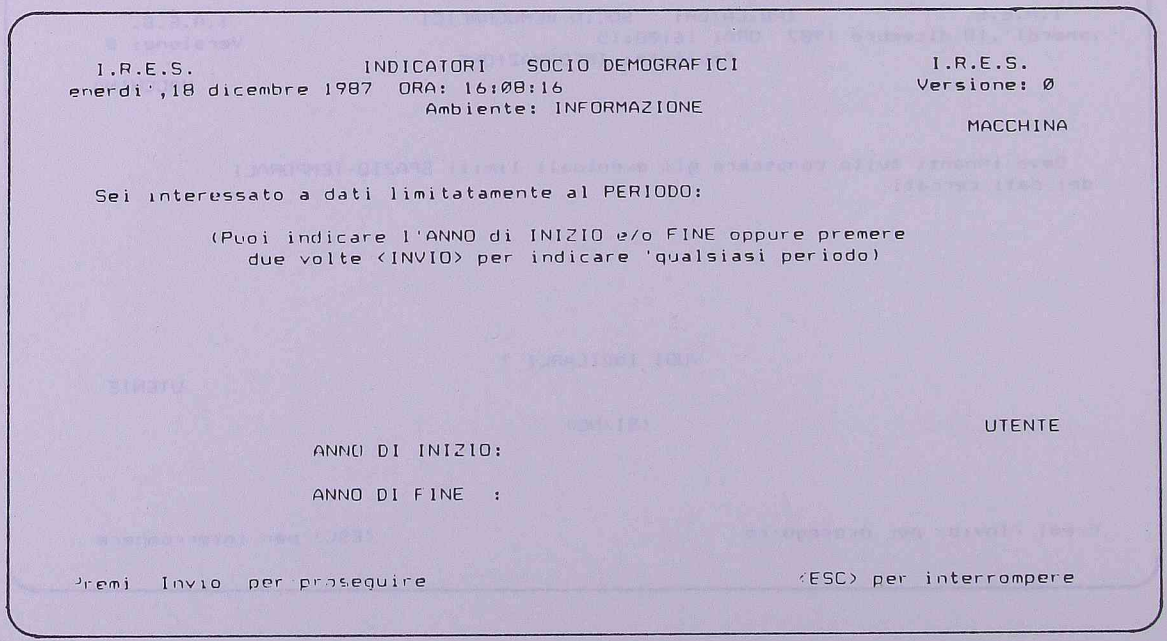

I.R.E.S. INDICATORI SOCIO DEMOGRAFICI I.R.E.S.
enerdi',18 dicembre 1987 ORA: 16:08:16 Versione: Ø
Ambiente: INFORMAZIONE
MACCHINA
Sei interessato a dati limitatamente al PERIODO:
(Puoi indicare l'ANNO di INIZIO e/o FINE oppure premere
due volte <INVIO> per indicare 'qualsiasi periodo)
UTENTE
ANNO DI INIZIO:
ANNO DI FINE :
Premi Invio per proseguire
<ESC> per interrompere

Fig. P4.1

```
I.R.E.S.                INDICATORI   SOCIO DEMOGRAFICI              I.R.E.S.
venerdi',18 dicembre 1987  ORA: 16:08:19                          Versione: Ø
                          Ambiente: INFORMAZIONE
                                                                    MACCHINA
   Posso verificare la disponibilità di dati sulla base dei seguenti
                        limiti SPAZIO-TEMPORALI:


   Limite spaziale: QUALSIASI LIVELLO TERRITORIALE

   Limite temporale: QUALSIASI PERIODO

                    CORRISPONDONO A QUELLI DESIDERATI ?
                                                                    UTENTE
                              (SI/NO)


Premi <Invio> per proseguire                          <ESC> per interrompere
```

Fig. P5

```
I.R.E.S.                INDICATORI   SOCIO DEMOGRAFICI              I.R.E.S.
venerdi',18 dicembre 1987  ORA: 16:08:24                          Versione: Ø
                          Ambiente: INFORMAZIONE
                                                                    MACCHINA

   Sto verificando la disponibilità dei dati richiesti.


                          PREGO ATTENDERE ...
```

Fig. P6

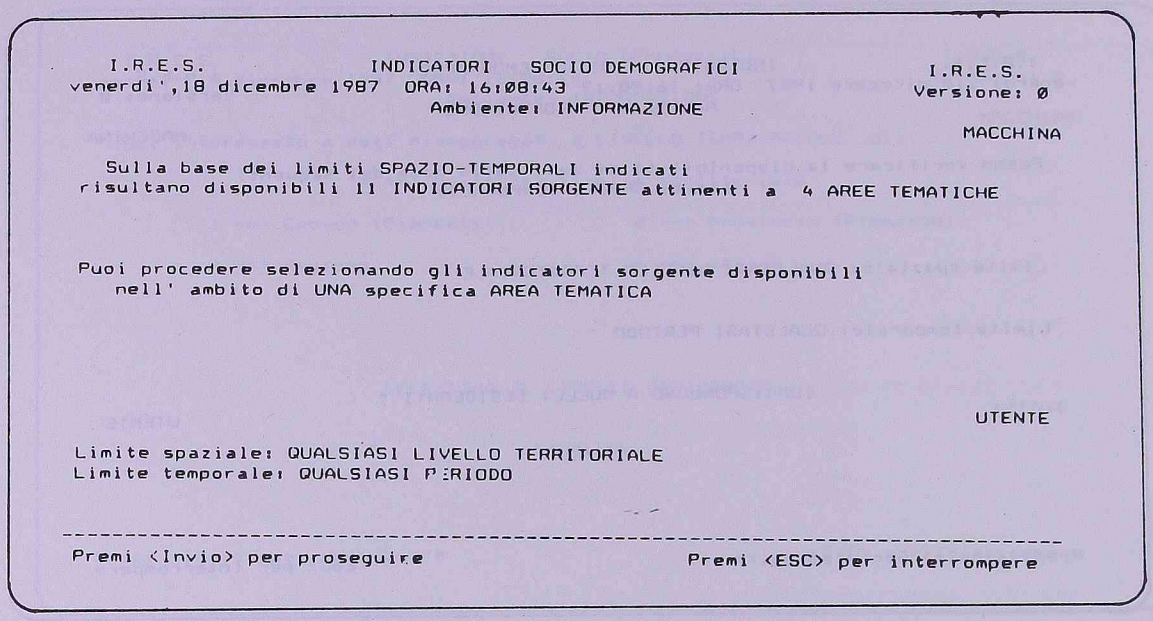


Fig. P6.1

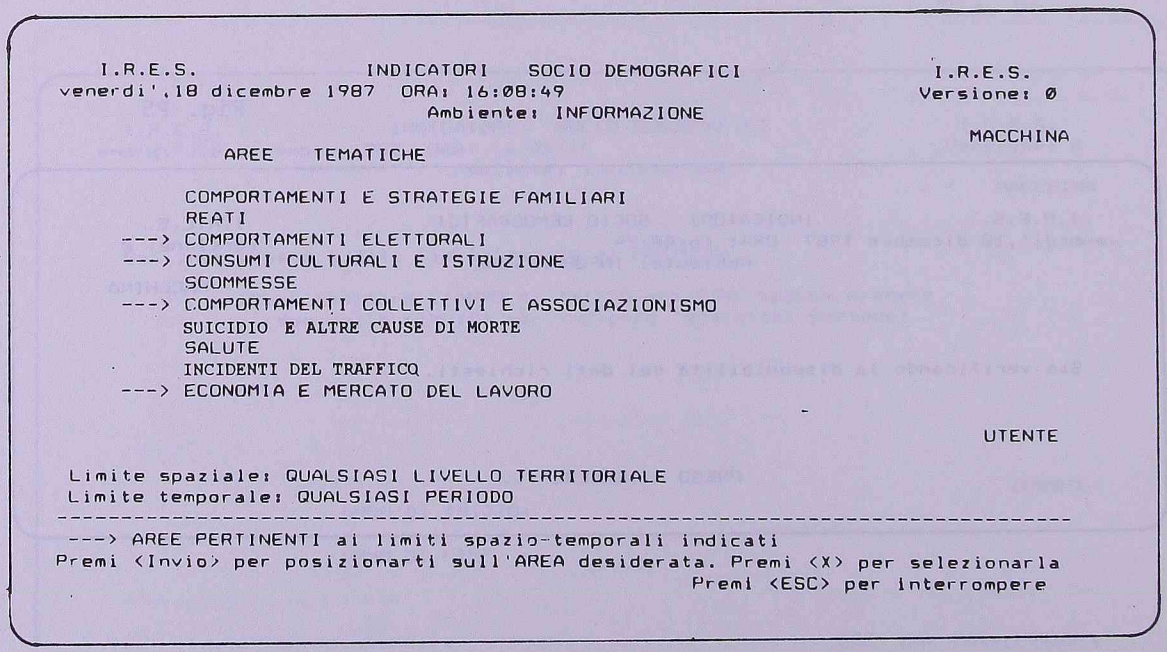


Fig. P6.2

```
I.R.E.S.                INDICATORI   SOCIO DEMOGRAFICI                I.R.E.S.
venerdi',18 dicembre 1987  ORA: 16:08:58                          Versione: Ø
                                Ambiente: INFORMAZIONE
                                                                      MACCHINA
          SOTTOAREE     dell' AREA TEMATICA: CONSUMI CULTURALI E ISTRUZIONE

     ---> Istruzione
          Partecipazione a spettacoli
          Partecipazione a manifestazioni sportive
     ---> Letture
          Ascolto radio e televisione
          Altro




                                                                      UTENTE
 Limite spaziale: QUALSIASI LIVELLO TERRITORIALE
 Limite temporale: QUALSIASI PERIODO
--------------------------------------------------------------------------------
 ---> SOTTOAREE PERTINENTI ai limiti spazio-temporali indicati
Premi <Invio> per posizionarti. <X> per selezionare la SOTTOAREA desiderata
                                                 Premi <ESC> per interrompere
```

Fig. P6.3

```
I.R.E.S.                INDICATORI   SOCIO DEMOGRAFICI                I.R.E.S.
venerdi',18 dicembre 1987  ORA: 16:09:10                          Versione: Ø
                                Ambiente: INFORMAZIONE
                                                                      MACCHINA
  Sulla base dei limiti SPAZIO-TEMPORALI e del TEMA indicati
risultano disponibili  3 INDICATORI SORGENTE


E' ora necessario selezionare quello desiderato ...




                                                                      UTENTE
Limite spaziale: QUALSIASI LIVELLO TERRITORIALE
Limite temporale: QUALSIASI PERIODO
--------------------------------------------------------------------------------

Premi <Invio> per prosegguire                          <ESC> per interrompere
```

Fig. P7

```
   I.R.E.S.              INDICATORI   SOCIO DEMOGRAFICI                I.R.E.S.
venerdi',18 dicembre 1987  ORA: 16:09:15                              Versione: Ø
                              Ambiente: INFORMAZIONE

                                                                       MACCHINA
    INDICATORE SORGENTE    (Variante)   FONTE       PERIODO  TERRITORIO LACUNE
   STUDENTI                      (Ø)   ISTAT      1954-1983 Prov.+Ita.   NO
   SCUOLE                        (Ø)   ISTAT      1954-1983 Prov.+Ita.   NO
   INSEGNANTI                    (Ø)   ISTAT      1954-1983 Prov.+Ita.   NO









                                                                  UTENTE


Premi <X> per selezionare.<H> per ottenere informazioni dettagliate

 Premi <Invio> per posizionarti                          <ESC> per interrompere
```

Fig. P7.1

```
   I.R.E.S.              INDICATORI   SOCIO DEMOGRAFICI                I.R.E.S.
venerdi',18 dicembre 1987  ORA: 16:10:23                              Versione: Ø
                              Ambiente: INFORMAZIONE
          Informazioni sull'INDICATORE SORGENTE   --- (Variante:  Ø)   MACCHINA
  STUDENTI

 AREA TEMATICA: CONSUMI CULTURALI E ISTRUZIONE
         SOTTOAREA: Istruzione
 FONTE: ISTAT                        ESTENSIONE: UNIVERSO
 SCALA: ASSOLUTA          UNITA' di CONTO O DI MISURA: PERSONE
 TIPO DI MISURAZIONE: PUNTUALE      LIVELLO TERRITORIALE: Prov.+Ita.
 CADENZA: ANNO                   PERIODO DI RIFERIMENTO: ANNO
 CONFINE TEMPORALE. DAL 1954 AL 1983 SENZA LACUNE
  DAL ANNO SC. 1954/55     FINO AL ANNO SC. 1983/84
 VETTORI PRIMARI DISPONIBILI:   464 SU 4 ELEMENTI DI DISAGGREGAZIONE

 COMMENTO: Si tratta di dati ricavati dagli annuari ISTAT registrati e organizza
ti dall'IRES secondo "TABELLE" la cui struttura è mantenuta invariata negli anni
.DESCRIZIONE OPERATIVA:




 Premi <Invio> per proseguire
```

Fig. P8

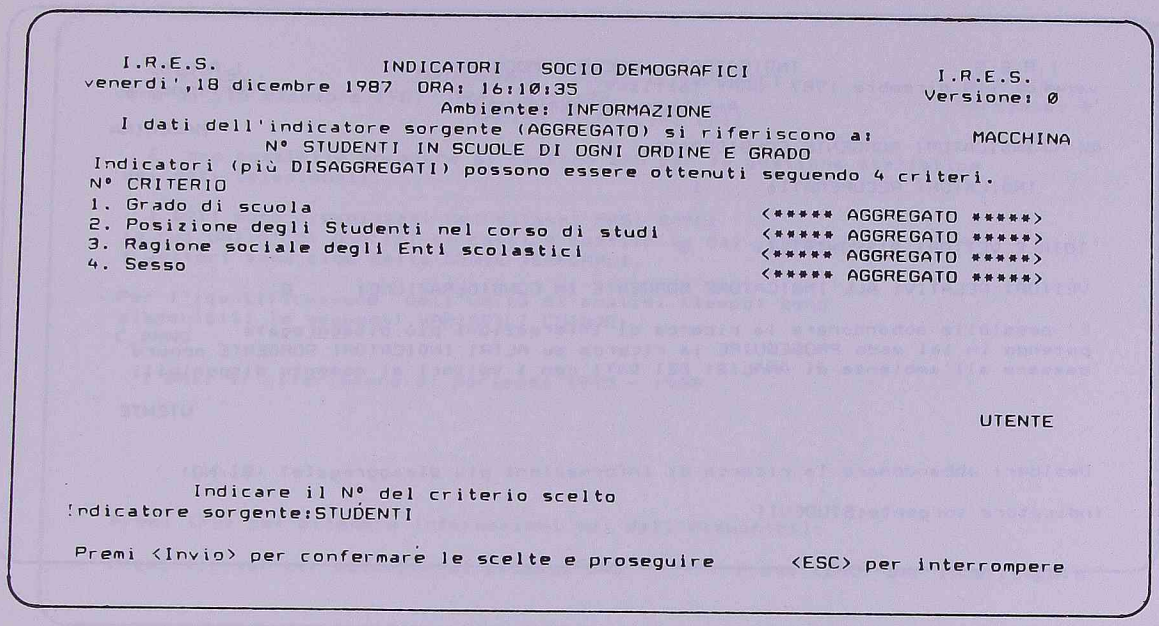


Fig. P8.1

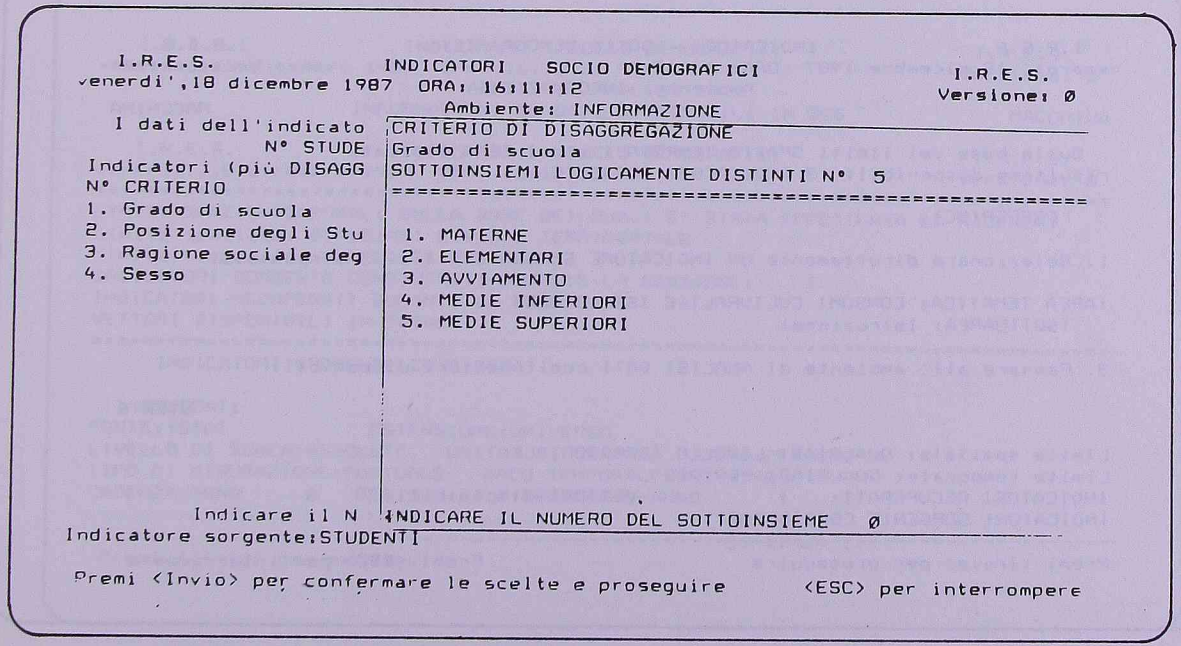


Fig. P9

```
   I.R.E.S.              INDICATORI   SOCIO DEMOGRAFICI              I.R.E.S.
venerdi',18 dicembre 1987  ORA: 16:11:49                          Versione: Ø
                              Ambiente: INFORMAZIONE
                                                                    MACCHINA
     INDICATORI SORGENTE CONSIDERATI:  1

     INDICATORI RECUPERATI:      1

 TOTALE VETTORI DISPONIBILI:      8

 VETTORI RELATIVI ALL'INDICATORE SORGENTE IN CONSIDERAZIONE:      8

 E' possibile abbandonare la ricerca di informazioni più disaggregate
 potendo in tal modo PROSEGUIRE la ricerca su ALTRI INDICATORI SORGENTE oppure
 passare all'ambiente di ANALISI DEI DATI con i vettori al momento disponibili

                                                                    UTENTE

 Desideri abbandonare la ricerca di informazioni più disaggregate? (SI/NO)

Indicatore sorgente:STUDENTI
```

Fig. P10

```
   I.R.E.S.              INDICATORI   SOCIO DEMOGRAFICI              I.R.E.S.
venerdi',18 dicembre 1987  ORA: 16:12:14                          Versione: Ø
                              Ambiente: INFORMAZIONE
                                                                    MACCHINA
   Sulla base dei limiti SPAZIO-TEMPORALI e del TEMA indicato
 risultano disponibili  3 INDICATORI SORGENTE

     PREFERISCI:

 1. Selezionare direttamente UN INDICATORE SORGENTE fra i  3 disponibili

 (AREA TEMATICA: CONSUMI CULTURALI E ISTRUZIONE)
    (SOTTOAREA: Istruzione)

 3. Passare all' ambiente di ANALISI DATI con i vettori disponibili

                                                                    UTENTE

 Limite spaziale: QUALSIASI LIVELLO TERRITORIALE
 Limite temporale: QUALSIASI PERIODO
 INDICATORI RECUPERATI:    1          VETTORI disponibili:      8
 INDICATORI SORGENTE CONSIDERATI: 1
-------------------------------------------------------------------------------
 Premi <Invio> per proseguire                  Premi <ESC> per interrompere
```

Fig. P11

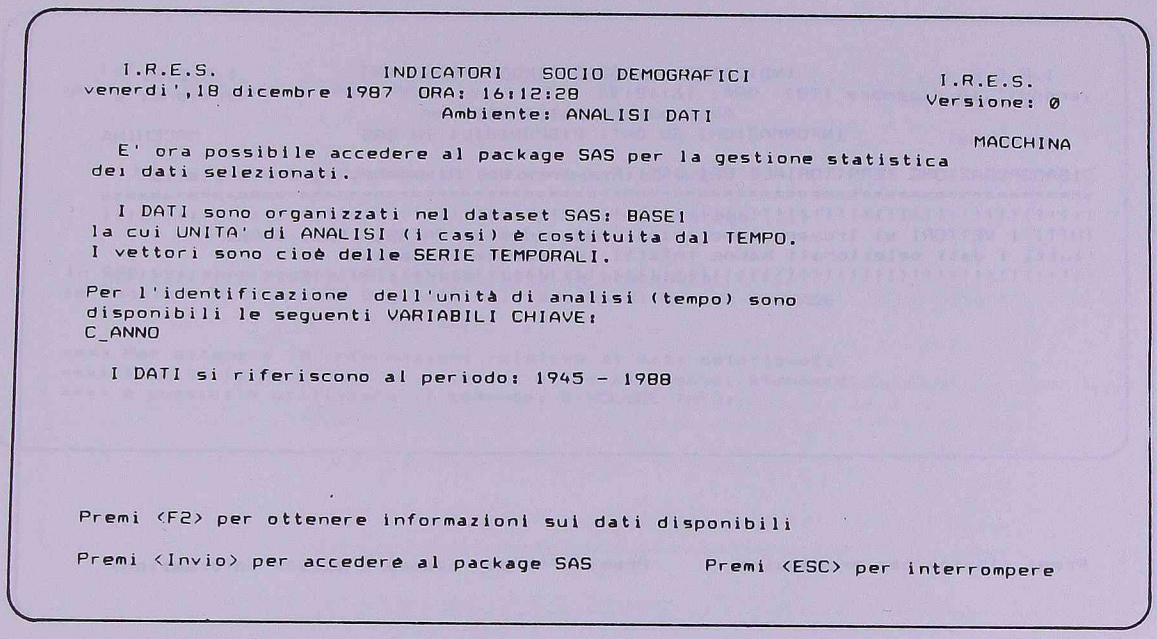


Fig. P11.1

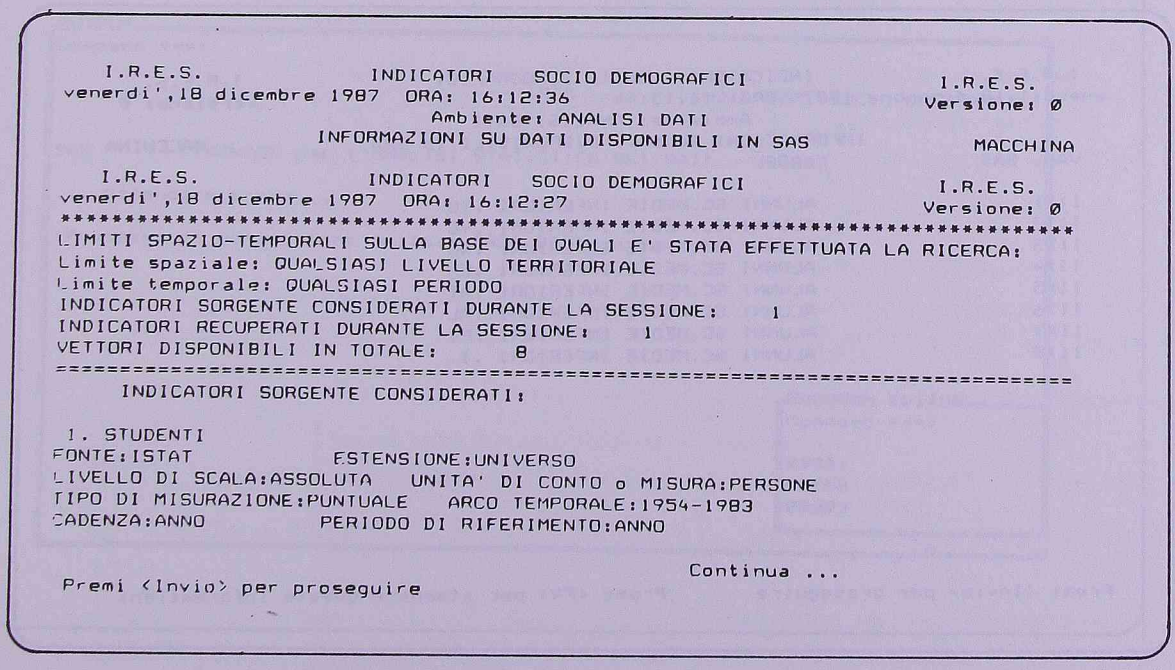


Fig. P11.2

```
   I.R.E.S.              INDICATORI   SOCIO DEMOGRAFICI              I.R.E.S.
venerdi',18 dicembre 1987  ORA: 16:12:36                            Versione: 0
                          Ambiente: ANALISI DATI
                    INFORMAZIONI SU DATI DISPONIBILI IN SAS              MACCHINA

DISAGGREGAZIONE TERRITORIALE DEI DATI:Province del Piemomte,Regione e Italia
================================================================================
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
TUTTI I VETTORI si trovano memorizzati nel medesimo DATASET SAS: BASE1
(tutti i dati selezionati hanno infatti la medesima cadenza)
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!


Premi <Invio> per proseguire       Premi <F4> per stampare queste informazioni
```

Fig. P11.3

```
   I.R.E.S.              INDICATORI   SOCIO DEMOGRAFICI              I.R.E.S.
venerdi',18 dicembre 1987  ORA: 16:13:44                            Versione: 0
                          Ambiente: ANALISI DATI
                    INFORMAZIONI SU DATI DISPONIBILI IN SAS              MACCHINA
   VAR. SAS              LABEL

   I1X1                  ALUNNI SC.MEDIE INFERIORI .TO
   I1X2                  ALUNNI SC.MEDIE INFERIORI .VC
   I1X3                  ALUNNI SC.MEDIE INFERIORI .NO
   I1X4                  ALUNNI SC.MEDIE INFERIORI .CN
   I1X5                  ALUNNI SC.MEDIE INFERIORI .AT
   I1X6                  ALUNNI SC.MEDIE INFERIORI .AL
   I1X7                  ALUNNI SC.MEDIE INFERIORI .P.
   I1X8                  ALUNNI SC.MEDIE INFERIORI .I.


Premi <Invio> per proseguire       Premi <F4> per stampare queste informazioni
```

Fig. P12

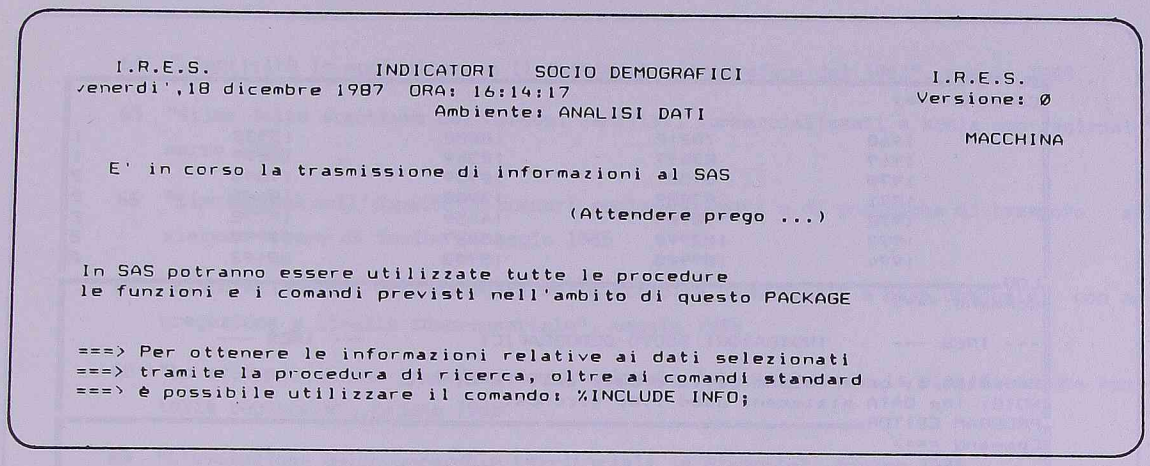


Fig. P12.1

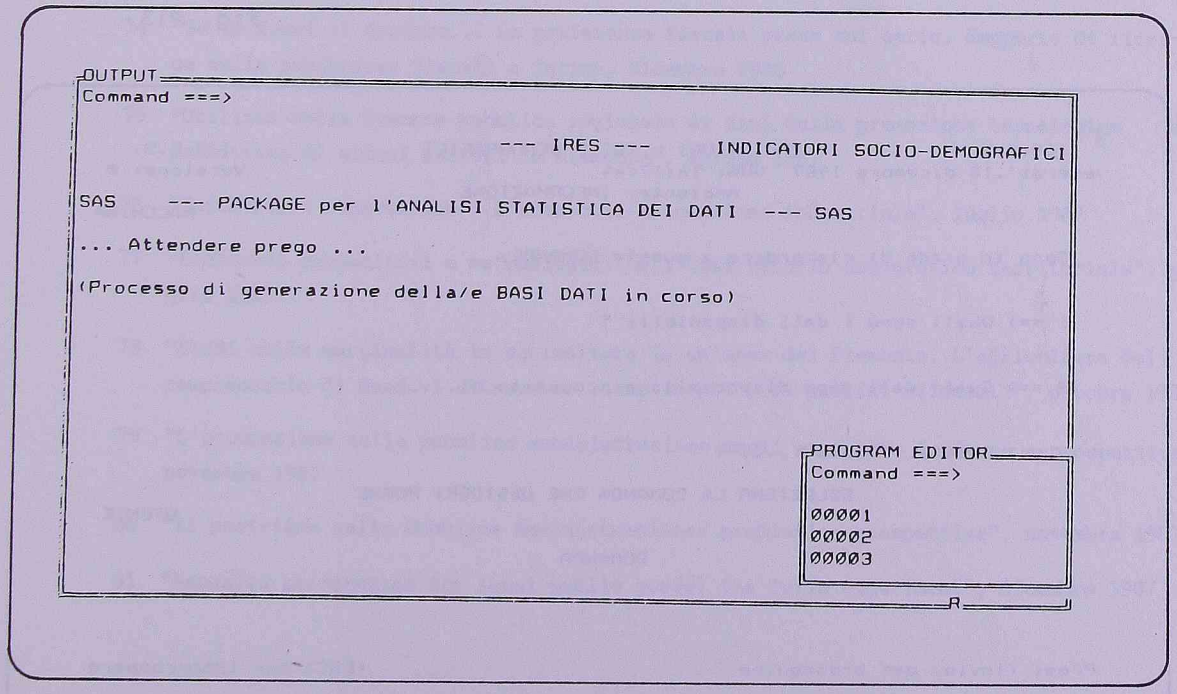


Fig. P12.2

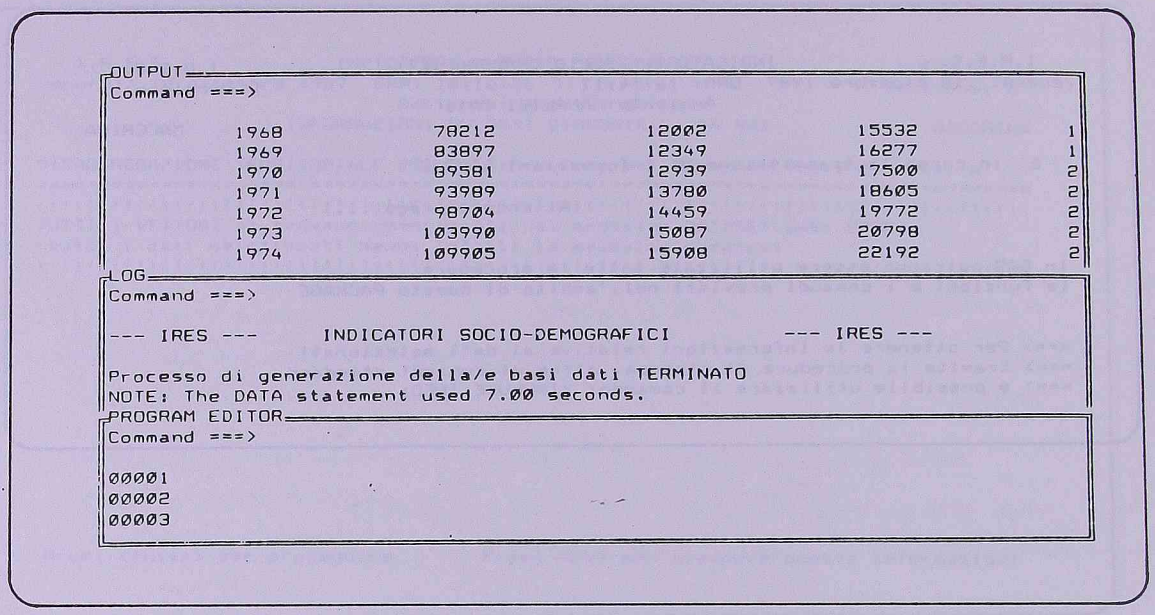
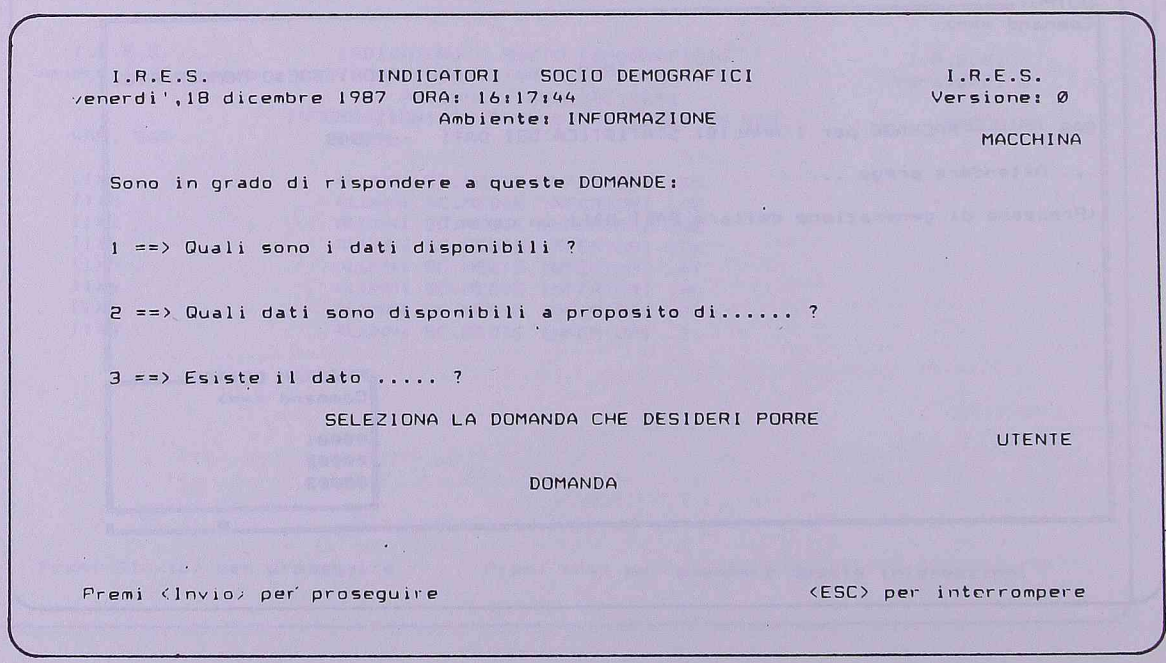


Fig. P13



ULTIMI WORKING PAPERS

64 "L'attività in agricoltura e il censimento demografico del 1981", maggio 1985

65 "Stima della struttura dei consumi familiari commercializzati a scala sub-regionale", marzo 1985

66 "Simulazione dell'impatto di scenari socio-economici e di politiche di trasporo sul sistema urbano di Torino", maggio 1985

67 "Elaborazione dei dati censuari sulle attività commerciali a base comunale, con aggregazione a livello comprensoriale", maggio 1985

68 "Lo sviluppo di una procedura computerizzata interattiva per la pianificazione sanitaria regionale", giugno 1985

69 "L'evoluzione delle gerarchie territoriali in Piemonte", giugno 1985

70 "An integrated model for the dynamic analisis of location-transport interrelation" , luglio 1985

71 "L'Agricoltura piemontese nel 1984 attraverso i dati dell'Osservatorio Contabile Regionale (O.C.R.), aprile 1986

72 "Livello e qualità della vita in Piemonte", aprile 1986

73 "Valutazione delle quote di mercato e dei livelli di modernizzazione del sistema distributivo alimentare per aree subregionali, dicembre 1986

74 "Se io fossi il Sindaco... Le preferenze fiscali prese sul serio. Rapporto di ricerca sulle preferenze fiscali a Torino, dicembre 1986

75 "Utilizzo della domanda pubblica regionale ai fini della promozione tecnologica e produttiva di alcuni settori in Piemonte", aprile 1987

76 "Industria e innovazione - L'area dell'automazione industriale", luglio 1987

77 "Elaborati conoscitivi e metodologici dell'Osservatorio demografico territoriale",luglio 1987

78 "Studi sulla marginalità in agricoltura in un'area del Piemonte. L'agricoltura del comprensorio di Mondovì attraverso i censimenti e le analisi aziendali", ottobre 1987

79 "L'occupazione nella pubblica amministrazione negli anni '80: tendenze e prospettive", novembre 1987

80 "Il part-time nella Pubblica Amministrazione: problemi e prospettive", novembre 1987

81 "Revealed preferences for local public goods: the Turin experiment", dicembre 1987

82 "Il problema dei flussi scolastici: un modello di analisi", dicembre 1987

83 "L'agricoltura a tempo parziale in Piemonte: un'analisi dei dati del III Censimento generale dell'agricoltura", marzo 1988

L'IRES è stato costituito nel 1958 dalla Provincia e dal Comune di Torino, con la partecipazione di altri enti pubblici e privati. Con la successiva adesione delle altre Province piemontesi, l'Istituto ha assunto carattere regionale.

Nel 1974 l'IRES è diventato ente strumentale della Regione Piemonte ed è stato dotato di personalità giuridica di diritto pubblico.

L'attività dell'IRES è attualmente disciplinata dalla legge regionale 18 febbraio 1985, n. 12.

L'IRES, struttura primaria di ricerca della Regione Piemonte, sviluppa la propria attività in raccordo con le esigenze della azione programmatoria ed operativa della Regione stessa, degli Enti locali e degli Enti pubblici.

Costituiscono oggetto dell'attività dell'Istituto:

- la redazione della relazione annuale sull'andamento socio-economico e territoriale della Regione;
- la conduzione di una permanente attività di osservazione, documentazione ed analisi sulle principali grandezze socio-economiche e territoriali del sistema regionale;
- lo svolgimento di periodiche rassegne congiunturali sull'economia regionale;
- lo svolgimento delle ricerche connesse alla redazione ed alla attuazione del piano regionale di sviluppo;
- lo svolgimento di ricerche di settore per conto della Regione e altri enti.

ires
ISTITUTO RICERCHE ECONOMICO - SOCIALI DEL PIEMONTE
VIA BOGINO 21 10123 TORINO